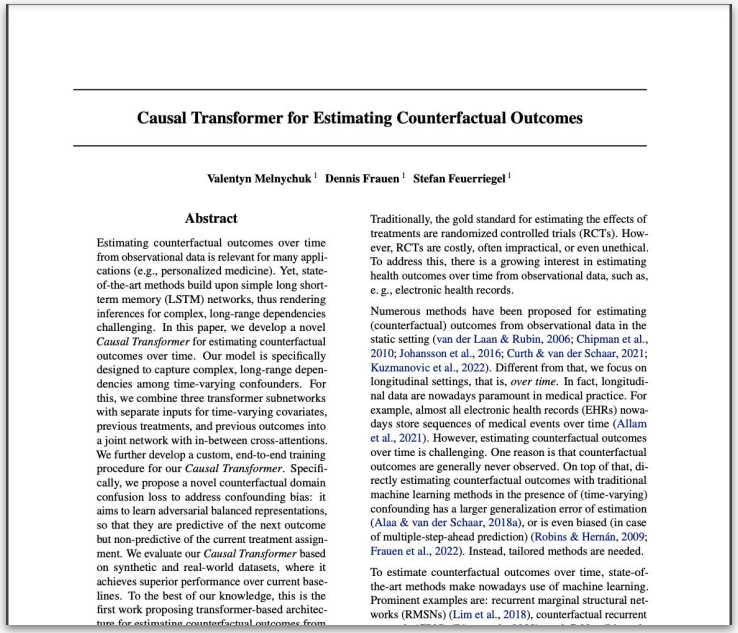
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT



Causal Transformer for Estimating Counterfactual Outcomes

Valentyn Melnychuk [1] Dennis Frauen [1] Stefan Feuerriegel [1]

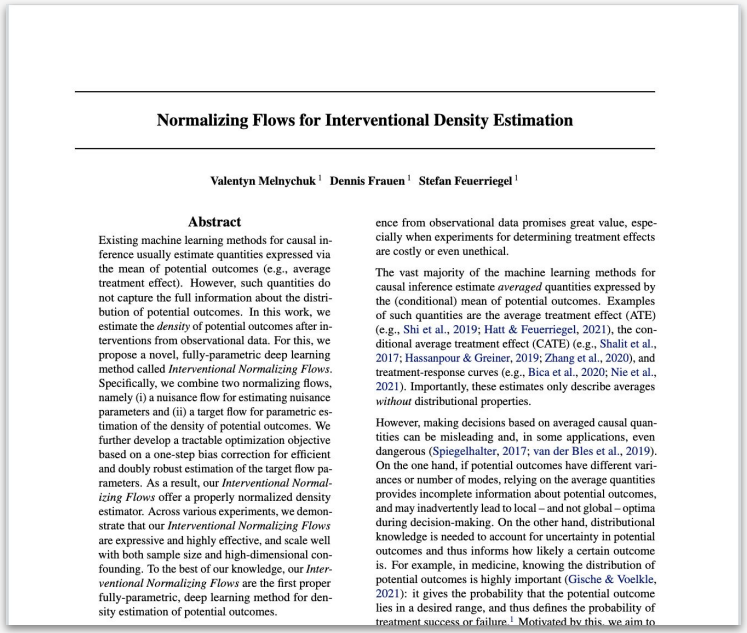
Abstract

Estimating counterfactual outcomes over time from observational data is relevant for many applications (e.g., personalized medicine). Yet, state-of-the-art methods build upon simple long short-term memory (LSTM) networks, thus rendering inferences for complex, long-range dependencies challenging. In this paper, we develop a novel *Causal Transformer* for estimating counterfactual outcomes over time. Our model is specifically designed to capture complex, long-range dependencies among time-varying confounders. For this, we combine three transformer subnetworks with separate inputs for time-varying covariates, previous treatments, and previous outcomes into a joint network with in-between cross-attentions. We further develop a custom, end-to-end training procedure for our *Causal Transformer*. Specifically, we propose a novel counterfactual domain confusion loss to address confounding bias: it aims to learn adversarial balanced representations, so that they are predictive of the next outcome but non-predictive of the current treatment assignment. We evaluate our *Causal Transformer* based on synthetic and real-world datasets, where it achieves superior performance over current baselines. To the best of our knowledge, this is the first work proposing transformer-based architectures for estimating counterfactual outcomes from

Traditionally, the gold standard for estimating the effects of treatments are randomized controlled trials (RCTs). However, RCTs are costly, often impractical, or even unethical. To address this, there is a growing interest in estimating health outcomes over time from observational data, such as, e. g., electronic health records.

Numerous methods have been proposed for estimating (counterfactual) outcomes from observational data in the static setting (van der Laan & Rubin, 2006; Chipman et al., 2010; Johansson et al., 2016; Curth & van der Schaar, 2021; Kuzmanovic et al., 2022). Different from that, we focus on longitudinal settings, that is, *over time*. In fact, longitudinal data are nowadays paramount in medical practice. For example, almost all electronic health records (EHRs) nowadays store sequences of medical events over time (Allam et al., 2021). However, estimating counterfactual outcomes over time is challenging. One reason is that counterfactual outcomes are generally never observed. On top of that, directly estimating counterfactual outcomes with traditional machine learning methods in the presence of (time-varying) confounding has a larger generalization error of estimation (Alaa & van der Schaar, 2018a), or is even biased (in case of multiple-step-ahead prediction) (Robins & Hernán, 2009; Frauen et al., 2022). Instead, tailored methods are needed.

To estimate counterfactual outcomes over time, state-of-the-art methods make nowadays use of machine learning. Prominent examples are: recurrent marginal structural networks (RMSNs) (Lim et al., 2018), counterfactual recurrent



Normalizing Flows for Interventional Density Estimation

Valentyn Melnychuk [1] Dennis Frauen [1] Stefan Feuerriegel [1]

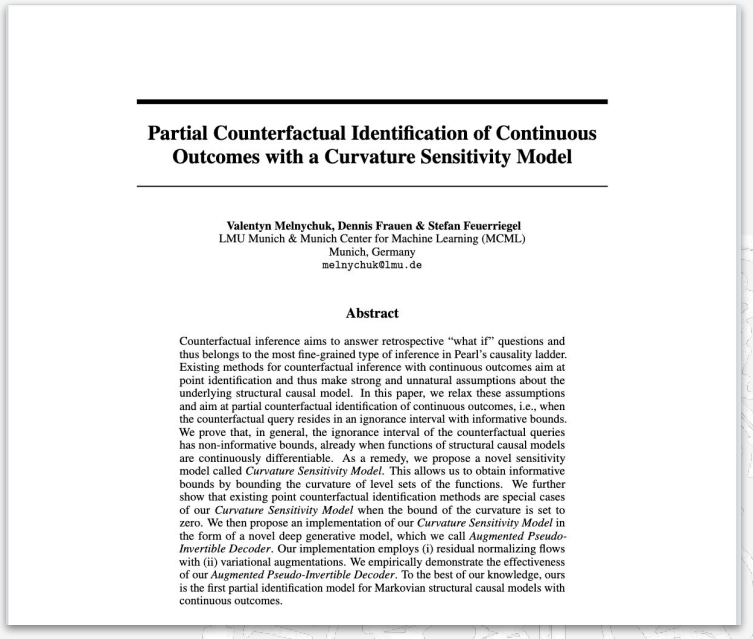
Abstract

Existing machine learning methods for causal inference usually estimate quantities expressed via the mean of potential outcomes (e.g., average treatment effect). However, such quantities do not capture the full information about the distribution of potential outcomes. In this work, we estimate the *density* of potential outcomes after interventions from observational data. For this, we propose a novel, fully-parametric deep learning method called *Interventional Normalizing Flows*. Specifically, we combine two normalizing flows, namely (i) a nuisance flow for estimating nuisance parameters and (ii) a target flow for parametric estimation of the density of potential outcomes. We further develop a tractable optimization objective based on a one-step bias correction for efficient and doubly robust estimation of the target flow parameters. As a result, our *Interventional Normalizing Flows* offer a properly normalized density estimator. Across various experiments, we demonstrate that our *Interventional Normalizing Flows* are expressive and highly effective, and scale well with both sample size and high-dimensional confounding. To the best of our knowledge, our *Interventional Normalizing Flows* are the first proper fully-parametric, deep learning method for density estimation of potential outcomes.

ence from observational data promises great value, especially when experiments for determining treatment effects are costly or even unethical.

The vast majority of the machine learning methods for causal inference estimate *averaged* quantities expressed by the (conditional) mean of potential outcomes. Examples of such quantities are the average treatment effect (ATE) (e.g., Shi et al., 2019; Hatt & Feuerriegel, 2021), the conditional average treatment effect (CATE) (e.g., Shalit et al., 2017; Hassanpour & Greiner, 2019; Zhang et al., 2020), and treatment-response curves (e.g., Bica et al., 2020; Nie et al., 2021). Importantly, these estimates only describe averages *without* distributional properties.

However, making decisions based on averaged causal quantities can be misleading and, in some applications, even dangerous (Spiegelhalter, 2017; van der Bles et al., 2019). On the one hand, if potential outcomes have different variances or number of modes, relying on the average quantities provides incomplete information about potential outcomes, and may inadvertently lead to local – and not global – optima during decision-making. On the other hand, distributional knowledge is needed to account for uncertainty in potential outcomes and thus informs how likely a certain outcome is. For example, in medicine, knowing the distribution of potential outcomes is highly important (Gische & Voelkle, 2021): it gives the probability that the potential outcome lies in a desired range, and thus defines the probability of treatment success or failure. [1] Motivated by this, we aim to



Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model

Valentyn Melnychuk, Dennis Frauen & Stefan Feuerriegel
LMU Munich & Munich Center for Machine Learning (MCML)
Munich, Germany
melnychuk@lmu.de

Abstract

Counterfactual inference aims to answer retrospective "what if" questions and thus belongs to the most fine-grained type of inference in Pearl's causality ladder. Existing methods for counterfactual inference with continuous outcomes aim at point identification and thus make strong and unnatural assumptions about the underlying structural causal model. In this paper, we relax these assumptions and aim at partial counterfactual identification of continuous outcomes, i.e., when the counterfactual query resides in an ignorance interval with informative bounds. We prove that, in general, the ignorance interval of the counterfactual queries has non-informative bounds, already when functions of structural causal models are continuously differentiable. As a remedy, we propose a novel sensitivity model called *Curvature Sensitivity Model*. This allows us to obtain informative bounds by bounding the curvature of level sets of the functions. We further show that existing point counterfactual identification methods are special cases of our *Curvature Sensitivity Model* when the bound of the curvature is set to zero. We then propose an implementation of our *Curvature Sensitivity Model* in the form of a novel deep generative model, which we call *Augmented Pseudo-Invertible Decoder*. Our implementation employs (i) residual normalizing flows with (ii) variational augmentations. We empirically demonstrate the effectiveness of our *Augmented Pseudo-Invertible Decoder*. To the best of our knowledge, ours is the first partial identification model for Markovian structural causal models with continuous outcomes.

PhD for Causal Machine Learning

Valentyn Melnychuk
LMU Munich + TUM, Munich, Germany
Workshop on Causal ML, 2023

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

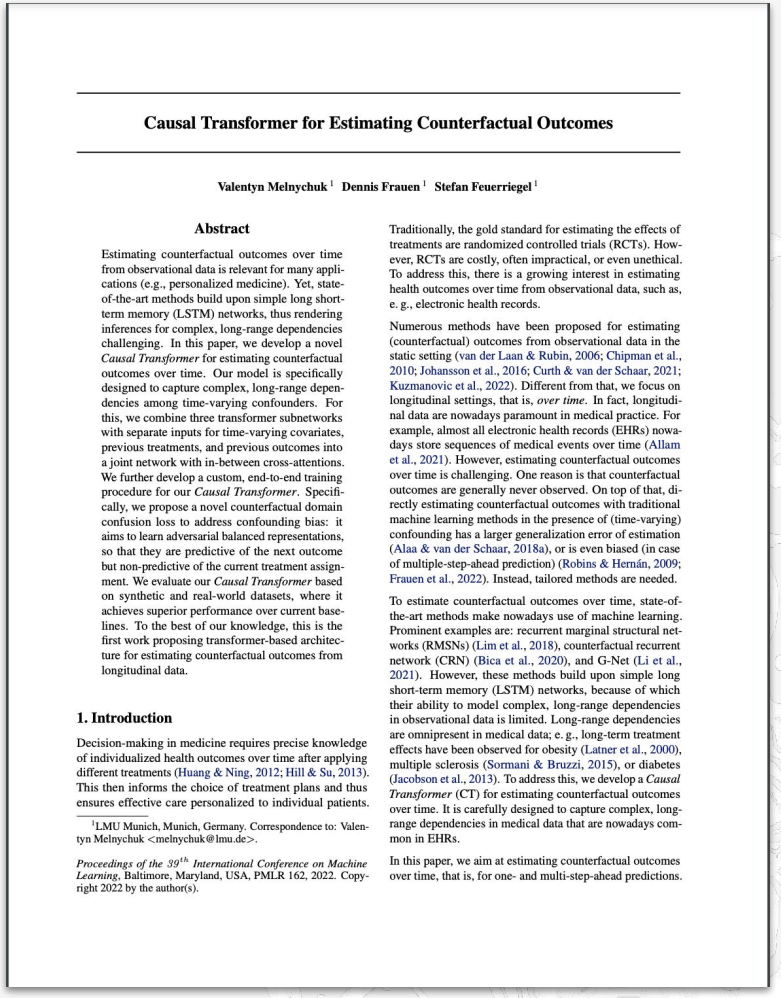
LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Causal Transformer for Estimating Counterfactual Outcomes

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

Accepted @ ICML 2022



Causal Transformer for Estimating Counterfactual Outcomes

Valentyn Melnychuk [1] Dennis Frauen [1] Stefan Feuerriegel [1]

Abstract

Estimating counterfactual outcomes over time from observational data is relevant for many applications (e.g., personalized medicine). Yet, state-of-the-art methods build upon simple long short-term memory (LSTM) networks, thus rendering inferences for complex, long-range dependencies challenging. In this paper, we develop a novel *Causal Transformer* for estimating counterfactual outcomes over time. Our model is specifically designed to capture complex, long-range dependencies among time-varying confounders. For this, we combine three transformer subnetworks with separate inputs for time-varying covariates, previous treatments, and previous outcomes into a joint network with in-between cross-attentions. We further develop a custom, end-to-end training procedure for our *Causal Transformer*. Specifically, we propose a novel counterfactual domain confusion loss to address confounding bias: it aims to learn adversarial balanced representations, so that they are predictive of the next outcome but non-predictive of the current treatment assignment. We evaluate our *Causal Transformer* based on synthetic and real-world datasets, where it achieves superior performance over current baselines. To the best of our knowledge, this is the first work proposing transformer-based architecture for estimating counterfactual outcomes from longitudinal data.

1. Introduction

Decision-making in medicine requires precise knowledge of individualized health outcomes over time after applying different treatments (Huang & Ning, 2012; Hill & Su, 2013). This then informs the choice of treatment plans and thus ensures effective care personalized to individual patients.

Traditionally, the gold standard for estimating the effects of treatments are randomized controlled trials (RCTs). However, RCTs are costly, often impractical, or even unethical. To address this, there is a growing interest in estimating health outcomes over time from observational data, such as, e. g., electronic health records.

Numerous methods have been proposed for estimating (counterfactual) outcomes from observational data in the static setting (van der Laan & Rubin, 2006; Chipman et al., 2010; Johansson et al., 2016; Curth & van der Schaar, 2021; Kuzmanovic et al., 2022). Different from that, we focus on longitudinal settings, that is, *over time*. In fact, longitudinal data are nowadays paramount in medical practice. For example, almost all electronic health records (EHRs) nowadays store sequences of medical events over time (Allam et al., 2021). However, estimating counterfactual outcomes over time is challenging. One reason is that counterfactual outcomes are generally never observed. On top of that, directly estimating counterfactual outcomes with traditional machine learning methods in the presence of (time-varying) confounding has a larger generalization error of estimation (Alaa & van der Schaar, 2018a), or is even biased (in case of multiple-step-ahead prediction) (Robins & Hernán, 2009; Frauen et al., 2022). Instead, tailored methods are needed.

To estimate counterfactual outcomes over time, state-of-the-art methods make nowadays use of machine learning. Prominent examples are: recurrent marginal structural networks (RMSNs) (Lim et al., 2018), counterfactual recurrent network (CRN) (Bica et al., 2020), and G-Net (Li et al., 2021). However, these methods build upon simple long short-term memory (LSTM) networks, because of which their ability to model complex, long-range dependencies in observational data is limited. Long-range dependencies are omnipresent in medical data; e. g., long-term treatment effects have been observed for obesity (Latner et al., 2000), multiple sclerosis (Sormani & Bruzzi, 2015), or diabetes (Jacobson et al., 2013). To address this, we develop a *Causal Transformer* (CT) for estimating counterfactual outcomes over time. It is carefully designed to capture complex, long-range dependencies in medical data that are nowadays common in EHRs.

In this paper, we aim at estimating counterfactual outcomes over time, that is, for one- and multi-step-ahead predictions.

[1]LMU Munich, Munich, Germany. Correspondence to: Valentyn Melnychuk <melnychuk@lmu.de>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Introduction: Estimating counterfactual outcomes over time

Why this is important?

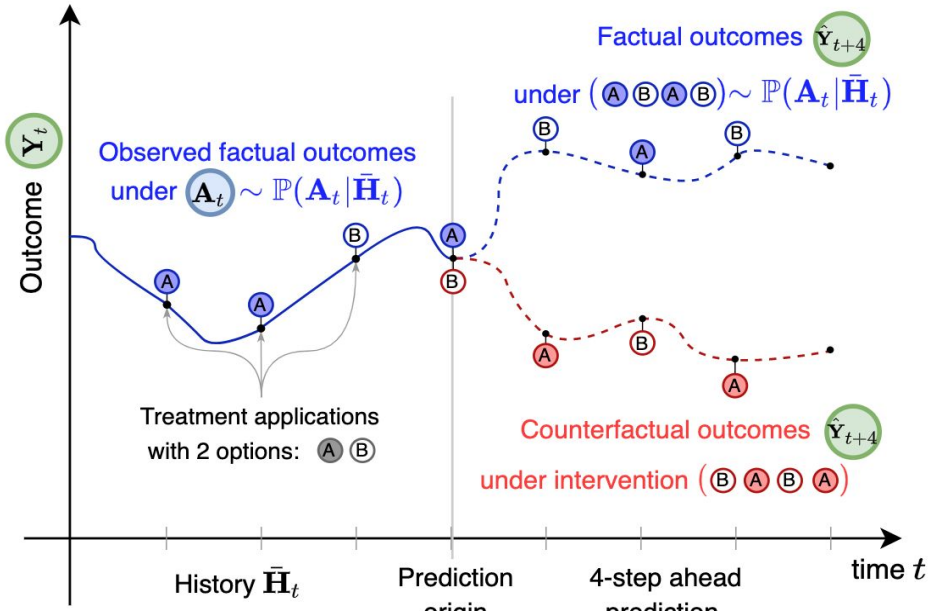
- Counterfactual prediction allows to answer **individualized** "what if" questions: what will happen to the patient, if I apply alternative sequence of treatments, **counterfactual**[1] to a standard treatment policy
- Growing opportunity to employ **observational data**:
  - randomized controlled trials (RCTs) are costly and/or unethical
  - abundance of large-scale observational data, e.g., electronic health records

Problem formulation

Given observational dataset of:

- $\mathbf{X}_t$ time-varying covariates (e.g., blood pressure)
- $\mathbf{V}$ static covariates (e.g., age)
- $\mathbf{A}_t$ treatments (e.g., ventilation)
- $\mathbf{Y}_t$ (factual[2]) outcomes (e.g., respiratory frequency)

we want to estimate **counterfactual outcomes over time** starting from prediction origin for a given sequence of treatment interventions



[1] Here, potential outcomes are meant, which correspond to the interventional level of valuation in Pearl's Hierarchy and the Foundations of Causal Inference

[2] Factual outcomes are observed under standard treatment policy

Introduction: Task complexity – Assumptions – Related methods

Why estimation is hard?

- Counterfactual outcomes are never directly observed in a real world
- Observed history grows with time
- Traditional machine learning is biased or sub-optimal in the presence of time-varying confounding[1]

Identifiability assumptions

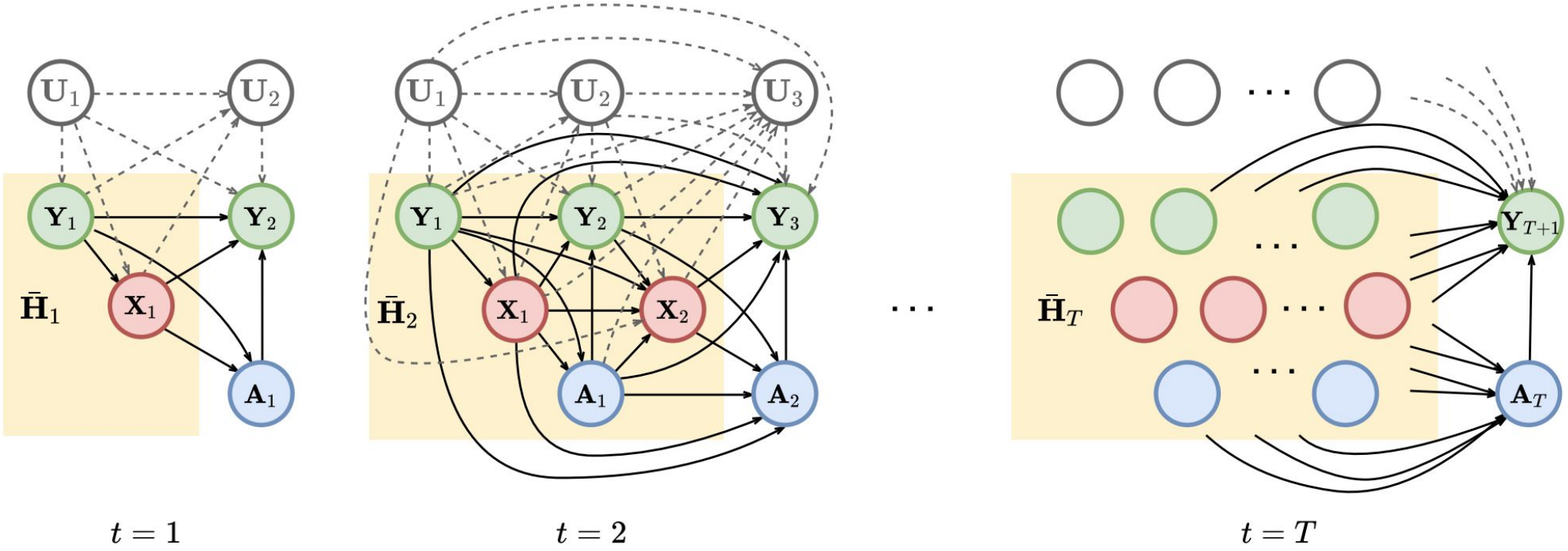
- **Consistency**. If $\bar{\mathbf{A}}_t = \bar{\mathbf{a}}_t$ is a given sequence of treatments for some patient, then $\mathbf{Y}_{t+1}[\bar{\mathbf{a}}_t] = \mathbf{Y}_{t+1}$.
- **Sequential Overlap.** There is always a non-zero probability of receiving/not receiving any treatment, conditioning on the previous history: $0 < \mathbb{P}(\mathbf{A}_t = \mathbf{a}_t \mid \bar{\mathbf{H}}_t = \bar{\mathbf{h}}_t) < 1$
- **Sequential Ignorability.** Current treatment is independent of the potential outcome, conditioning on the observed history $\mathbf{A}_t \perp\!\!\!\perp \mathbf{Y}_{t+1}[\mathbf{a}_t] \mid \bar{\mathbf{H}}_t$

Related methods

- **Marginal Structural Models (MSMs)** (Robins et al., 2000; Hernan et al., 2001): only linear modelling
- **Recurrent Marginal Structural Networks** (RMSNs) (Lim et al., 2018): several LSTM networks for inverse probability of treatment weights (IPTW) and prediction
- **Counterfactual Recurrent Network** (CRN) (Bica et al., 2020): encoder-decoder LSTMS with adversarial learning of treatment invariant representations
- **G-Net** (Li et al., 2021): G-computation on top of LSTM

[1] Time-varying confounding stands for a non-randomized treatment assignment, which depends on time-varying covariates, previous treatments and previous outcomes

Introduction: Causal diagram & Causal query



$$\mathbb{E}\left(\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}] \mid \bar{\mathbf{H}}_t\right)$$

- τ-step-ahead counterfactual outcome
- sequence of treatment interventions
- history before prediction origin

(individualized) expected counterfactual outcomes over time

Introduction: Research gap – Our contributions

Research gap

- Current state-of-the-art methods are built on top of long short-term memory (LSTM), thus rendering inferences for complex, long-range dependencies challenging

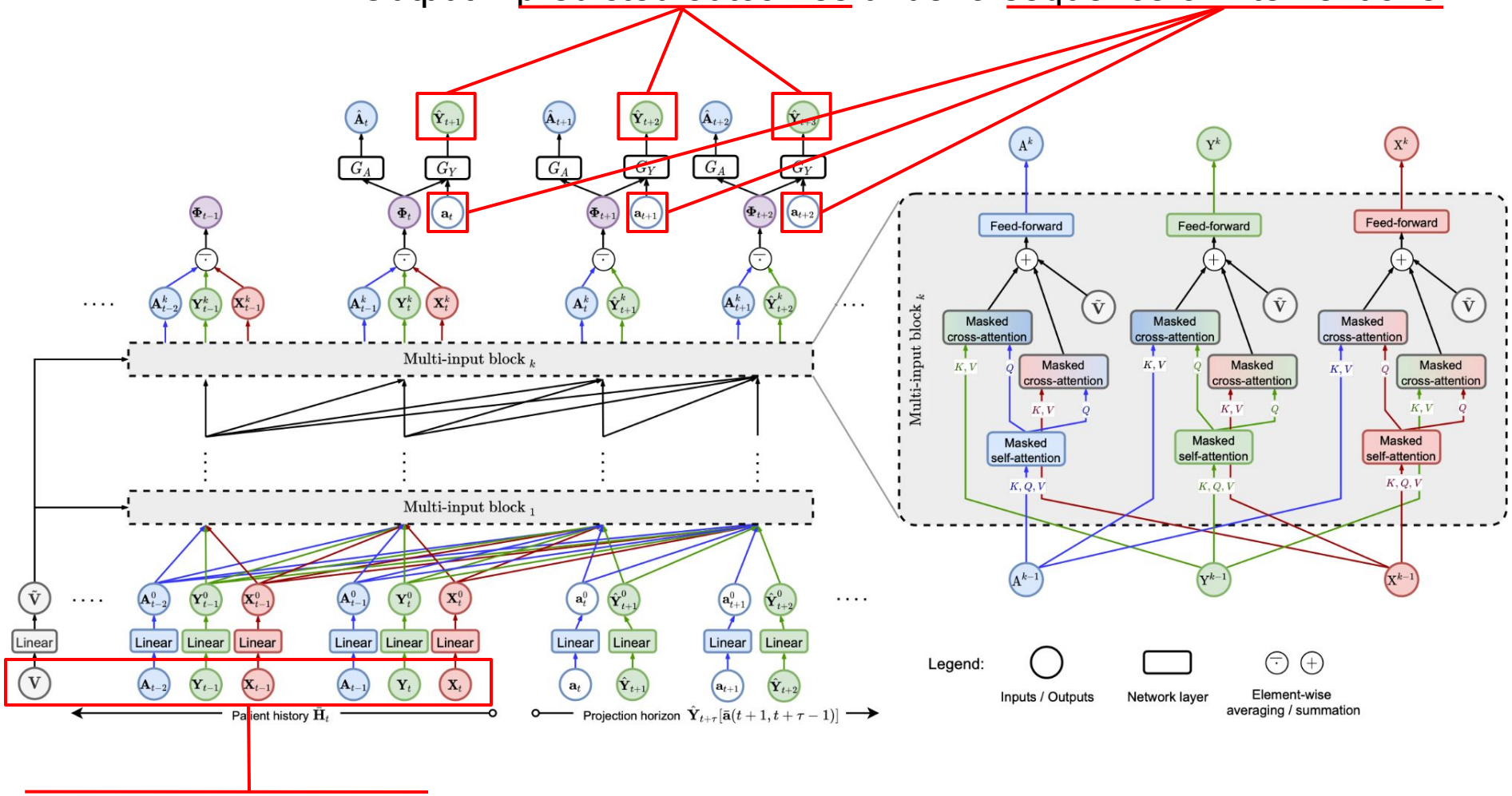
Our contributions

Causal Transformer (CT) is an end-to-end model, first tailoring of transformers to a counterfactual prediction task over time:

- CT captures **complex, long-range dependencies** between time-varying covariates, treatments and outcomes
- CT employs a novel **counterfactual domain confusion (CDC) loss** to address a time-varying confounding
- CT achieves **state-of-the-art performance** on synthetic, semi-synthetic & real benchmarks

Causal Transformer: Novel architecture

2. Output – predicted outcomes under a sequence of interventions

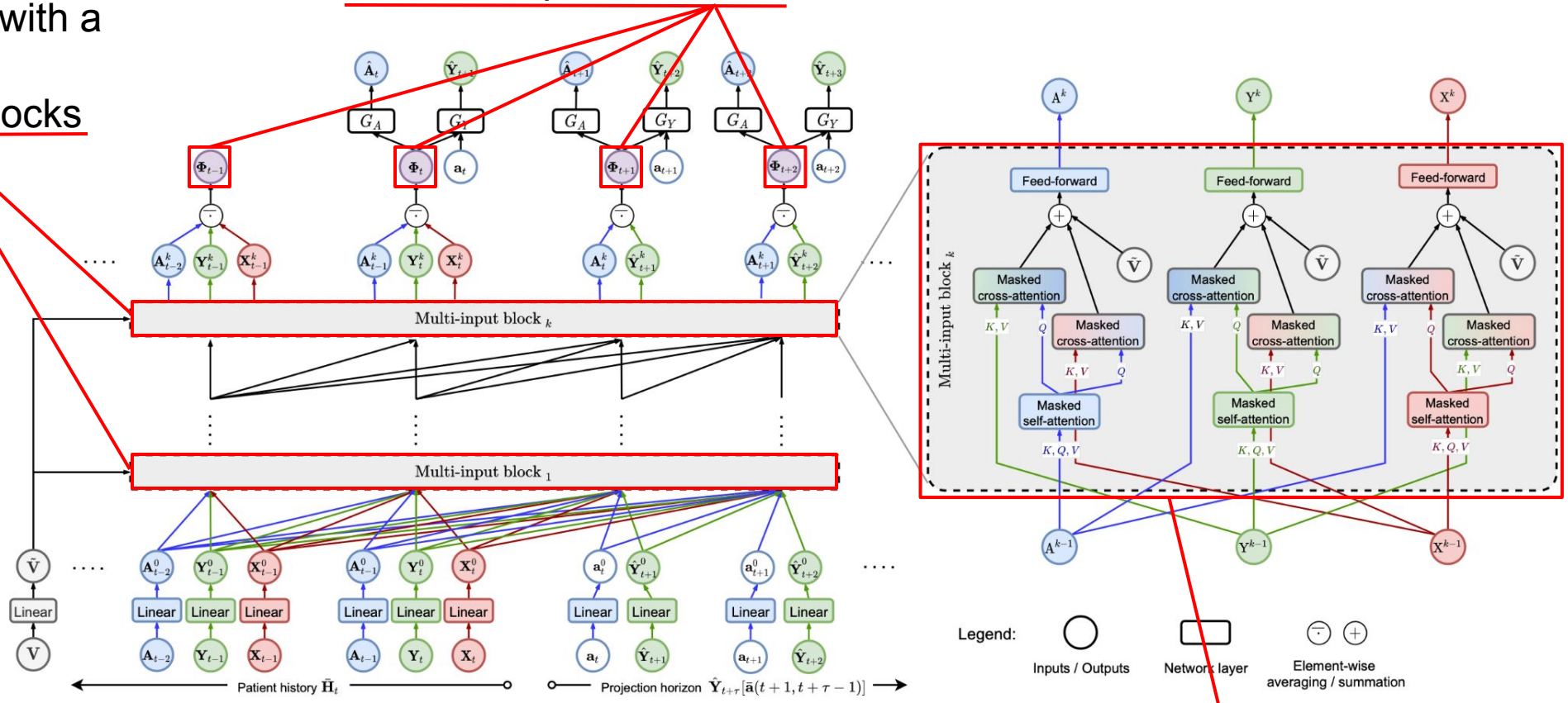


1. Input – observed patient history

Causal Transformer: Novel architecture

3. Inputs are transformed with a stack of multi-input blocks

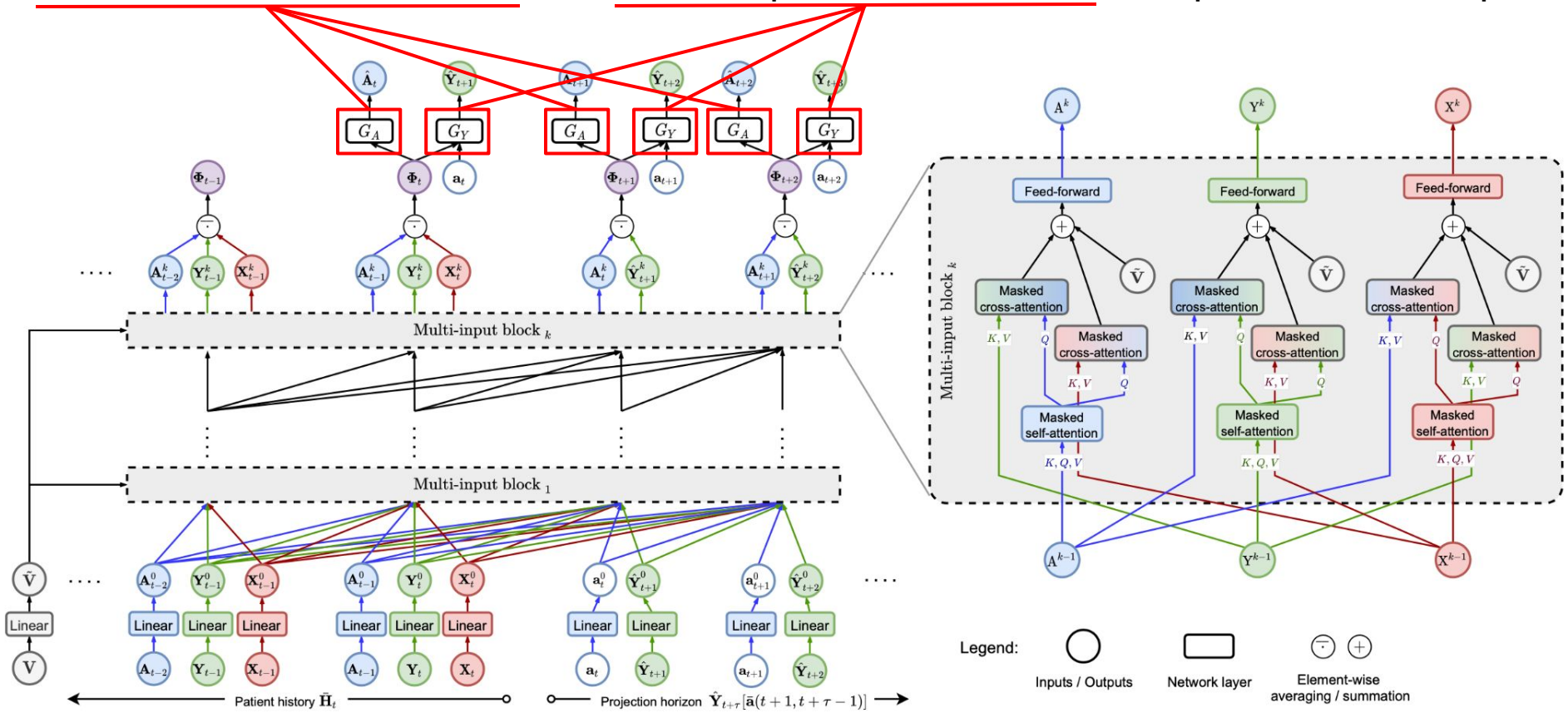
4. Outputs of the last block are averaged and form balanced representations



5. Each block is equipped with self-attention, cross-attention and feed-forward layers

Causal Transformer: Novel architecture

6. We place treatment classifier network and outcome prediction network on top of balanced representations



7. Both treatment classifier and outcome prediction networks are used for the novel counterfactual domain confusion loss (CDC) loss

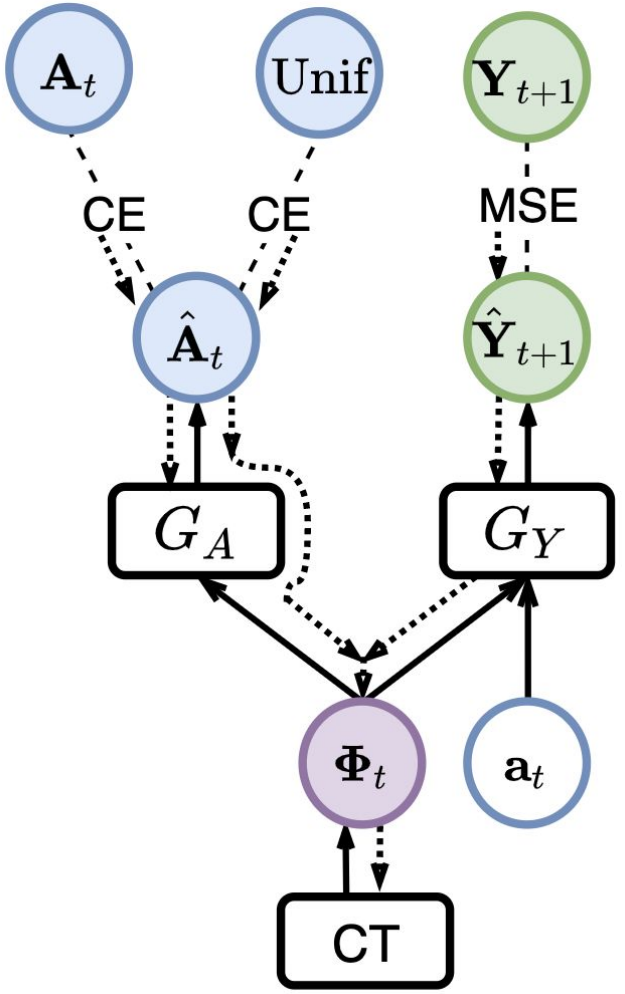
Causal Transformer: Counterfactual domain confusion (CDC) loss

- Idea stems from the unsupervised domain adaptation[1]
- CDC is an adversarial objective, which aims to

  (a) make **balanced representations** Φ_t non-predictive of the current treatment:

  - minimizing cross-entropy of current treatment wrt. G_A
  - minimizing cross-entropy between uniform treatment and output of treatment classifier network wrt. CT

  (b) at same time, make them **predictive of the outcome** wrt. CT and G_Y by minimizing factual MSE
- Adversarial learning is stabilized with exponential moving average (EMA of model weights



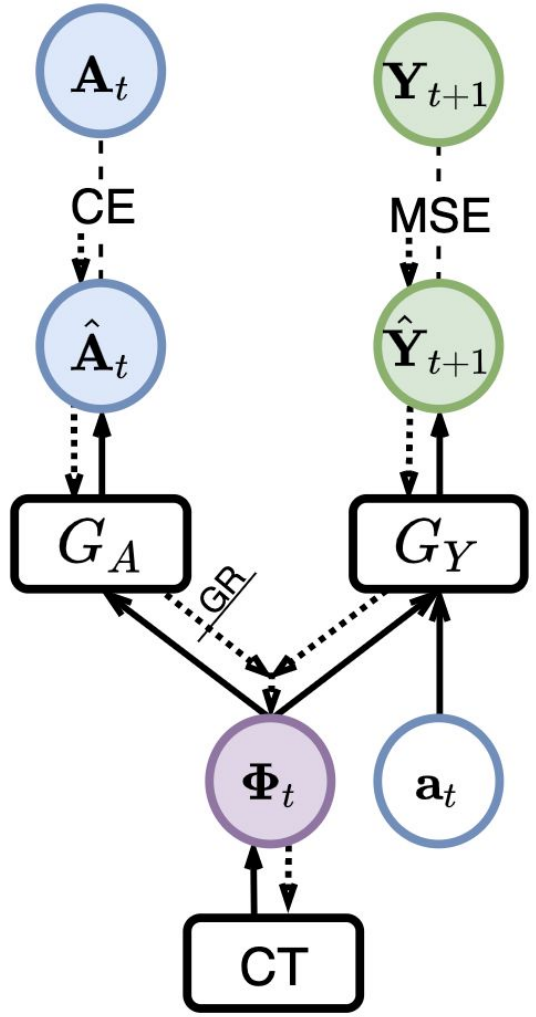
[1] Tzeng, Eric, et al. "Simultaneous deep transfer across domains and tasks." Proceedings of the IEEE international conference on computer vision (2015)

Causal Transformer: Theoretical insights

- Previously proposed **Gradient reversal**[1] (CRN, Bica et al., 2020) extends in two ways
- We prove a theorem, similar to (CRN, Bica et al., 2020): finding a solution to an adversarial objective of CDC loss renders distributions of representations conditional on each treatment **equal** (= balanced)
- In our case, we minimize a reversed KL-divergence:

CDC loss (our paper)	Gradient reversal (CRN, Bica et al., 2020)
Minimizing $\sum_{j=1}^{K} KL\left(\frac{1}{K}\sum_{i=1}^{K} P_i^{\Phi}(x') \middle\| P_j^{\Phi}(x')\right)$	Minimizing $\sum_{j=1}^{K} KL\left(P_j^{\Phi}(x') \middle\| \frac{1}{K}\sum_{i=1}^{K} P_i^{\Phi}(x')\right)$

where $P_j^{\Phi}(x')$ is a distribution of representation conditional on treatment j



[1] Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." International conference on machine learning. PMLR, 2015

Experiments: Datasets – Results

Datasets

- We evaluate CT based on synthetic, (self-designed) semi-synthetic and real-world (MIMIC-III) datasets
- Only synthetic and semi-synthetic data have ground-truth counterfactuals; real-world evaluation is a proof of concept
- We compared root-mean-squared error (RMSE) of one and multiple-step-ahead predictions

Results

CT achieves **superior performance** over current baselines for benchmarks with long-range dependencies and long prediction horizons, e.g., for semi-synthetic benchmark:

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
MSMs (Robins et al., 2000)	0.37 ± 0.01	0.57 ± 0.03	0.74 ± 0.06	0.88 ± 0.03	1.14 ± 0.10	1.95 ± 1.48	3.44 ± 4.57	> 10.0	> 10.0	> 10.0
RMSNs (Lim et al., 2018)	0.24 ± 0.01	0.47 ± 0.01	0.60 ± 0.01	0.70 ± 0.02	0.78 ± 0.04	0.84 ± 0.05	0.89 ± 0.06	0.94 ± 0.08	0.97 ± 0.09	1.00 ± 0.11
CRN (Bica et al., 2020)	0.30 ± 0.01	0.48 ± 0.02	0.59 ± 0.02	0.65 ± 0.02	0.68 ± 0.02	0.71 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.78 ± 0.02
G-Net (Li et al., 2021)	0.34 ± 0.01	0.67 ± 0.03	0.83 ± 0.04	0.94 ± 0.04	1.03 ± 0.05	1.10 ± 0.05	1.16 ± 0.05	1.21 ± 0.06	1.25 ± 0.06	1.29 ± 0.06
EDCT w/ GR ($\lambda = 1$) (*ours*)	0.29 ± 0.01	0.46 ± 0.01	0.56 ± 0.01	0.62 ± 0.01	0.67 ± 0.01	0.70 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.78 ± 0.01
CT ($\alpha = 0$) (*ours*) *	**0.20 ± 0.01**	**0.38 ± 0.01**	**0.45 ± 0.01**	0.50 ± 0.02	**0.52 ± 0.02**	0.55 ± 0.02	0.56 ± 0.02	0.58 ± 0.02	0.60 ± 0.02	0.61 ± 0.02
CT (*ours*)	**0.20 ± 0.01**	**0.38 ± 0.01**	**0.45 ± 0.01**	**0.49 ± 0.01**	**0.52 ± 0.02**	**0.53 ± 0.02**	**0.55 ± 0.02**	**0.56 ± 0.02**	**0.58 ± 0.02**	**0.59 ± 0.02**

Lower = better (best in bold)

Experiments: Ablation study

Ablation types

Based on synthetic datasets we evaluate different versions of CT with varying:

(a) different components within the subnetworks (positional encodings, attentional dropout)

(b) different losses (CDC vs Gradient reversal vs no balancing, w/ vs w/o EMA of weights)

(c) single-subnetwork variant of CT vs original CT

Results

- Combination of **end-to-end three subnetworks architecture and the novel CDC loss** is crucial (neither work better alone)
- Switching the backbone from LSTM to transformer and using gradient reversal as in CRN (Bica et al., 2020) gives worse results

		$\tau = 1$		$\tau = 6$	
		$\gamma = 1$	$\gamma = 4$	$\gamma = 1$	$\gamma = 4$
CT (proposed)		0.80	1.32	0.63	0.93
a	w/ non-trainable PE*	$\pm$0.00	$-$0.02	+0.01	$-$0.03
	w/ absolute PE*	+0.04	+0.16	+0.15	+1.00
	w/o attentional dropout*	$\pm$0.00	+0.07	+0.00	+0.09
	w/o cross-attention*	+0.03	+0.16	+0.06	+0.10
b	w/o EMA ($\beta = 0$)*	+0.03	+0.38	+0.03	+0.33
	w/o balancing ($\alpha = 0$; $\beta = 0.99$)*	$-$0.01	$-$0.02	$\pm$0.00	+0.07
	w/ GR ($\lambda = 1$)	+0.02	+0.17	+0.08	+0.33
c	EDCT w/ GR ($\lambda = 1$)	+0.16	+0.08	+0.05	+0.23
	EDCT w/ DC ($\alpha = 0.01$; $\beta = 0.99$)	$-$0.03	+0.10	$-$0.03	+0.23

Lower = better;

Open questions / Future work

- Multi-step-ahead prediction is **biased** with balanced representations. Theory of bias-variance tradeoff of the G-computation or IPW methods is missing for (individualized) expected counterfactual outcomes over time.

$$\mathbb{E}\left(\mathbf{Y}_{t+\tau}[\bar{\mathbf{a}}_{t:t+\tau-1}] \mid \bar{\mathbf{H}}_t\right) = \int_{\mathbb{R}^{d_x} \times \cdots \times \mathbb{R}^{d_x}} \mathbb{E}\left(\mathbf{Y}_{t+\tau} \mid \bar{\mathbf{H}}_t, \bar{\mathbf{x}}_{t+1:t+\tau-1}, \bar{\mathbf{y}}_{t+1:t+\tau-1}, \bar{\mathbf{a}}_{t:t+\tau-1}\right) \times$$

$$\prod_{j=t+1}^{t+\tau-1} \mathbb{P}\left(\mathbf{x}_j \mathbf{y}_j \mid \bar{\mathbf{H}}_t, \bar{\mathbf{x}}_{t+1:j-1}, \bar{\mathbf{y}}_{t+1:j-1}, \bar{\mathbf{a}}_{t:j-1}\right) \, \mathrm{d}\bar{\mathbf{x}}_{t+1:t+\tau-1} \, \mathrm{d}\bar{\mathbf{y}}_{t+1:t+\tau-1}$$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

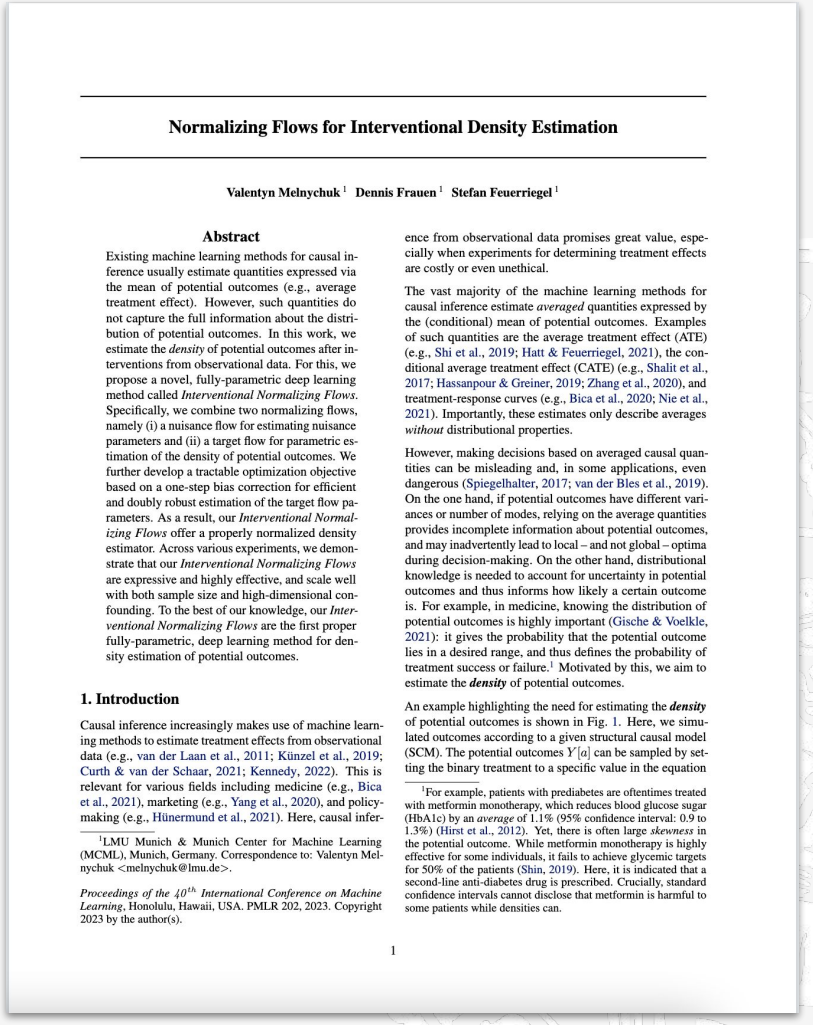
LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Normalizing Flows for Interventional Density Estimation

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

Accepted @ ICML 2023



Normalizing Flows for Interventional Density Estimation

Valentyn Melnychuk [1] Dennis Frauen [1] Stefan Feuerriegel [1]

Abstract

Existing machine learning methods for causal inference usually estimate quantities expressed via the mean of potential outcomes (e.g., average treatment effect). However, such quantities do not capture the full information about the distribution of potential outcomes. In this work, we estimate the *density* of potential outcomes after interventions from observational data. For this, we propose a novel, fully-parametric deep learning method called *Interventional Normalizing Flows*. Specifically, we combine two normalizing flows, namely (i) a nuisance flow for estimating nuisance parameters and (ii) a target flow for parametric estimation of the density of potential outcomes. We further develop a tractable optimization objective based on a one-step bias correction for efficient and doubly robust estimation of the target flow parameters. As a result, our *Interventional Normalizing Flows* offer a properly normalized density estimator. Across various experiments, we demonstrate that our *Interventional Normalizing Flows* are expressive and highly effective, and scale well with both sample size and high-dimensional confounding. To the best of our knowledge, our *Interventional Normalizing Flows* are the first proper fully-parametric, deep learning method for density estimation of potential outcomes.

1. Introduction

Causal inference increasingly makes use of machine learning methods to estimate treatment effects from observational data (e.g., van der Laan et al., 2011; Künzel et al., 2019; Curth & van der Schaar, 2021; Kennedy, 2022). This is relevant for various fields including medicine (e.g., Bica et al., 2021), marketing (e.g., Yang et al., 2020), and policy-making (e.g., Hünermund et al., 2021). Here, causal inference from observational data promises great value, especially when experiments for determining treatment effects are costly or even unethical.

The vast majority of the machine learning methods for causal inference estimate *averaged* quantities expressed by the (conditional) mean of potential outcomes. Examples of such quantities are the average treatment effect (ATE) (e.g., Shi et al., 2019; Hatt & Feuerriegel, 2021), the conditional average treatment effect (CATE) (e.g., Shalit et al., 2017; Hassanpour & Greiner, 2019; Zhang et al., 2020), and treatment-response curves (e.g., Bica et al., 2020; Nie et al., 2021). Importantly, these estimates only describe averages *without* distributional properties.

However, making decisions based on averaged causal quantities can be misleading and, in some applications, even dangerous (Spiegelhalter, 2017; van der Bles et al., 2019). On the one hand, if potential outcomes have different variances or number of modes, relying on the average quantities provides incomplete information about potential outcomes, and may inadvertently lead to local – and not global – optima during decision-making. On the other hand, distributional knowledge is needed to account for uncertainty in potential outcomes and thus informs how likely a certain outcome is. For example, in medicine, knowing the distribution of potential outcomes is highly important (Gische & Voelkle, 2021): it gives the probability that the potential outcome lies in a desired range, and thus defines the probability of treatment success or failure.[1] Motivated by this, we aim to estimate the ***density*** of potential outcomes.

An example highlighting the need for estimating the ***density*** of potential outcomes is shown in Fig. 1. Here, we simulated outcomes according to a given structural causal model (SCM). The potential outcomes $Y[a]$ can be sampled by setting the binary treatment to a specific value in the equation

[1]LMU Munich & Munich Center for Machine Learning (MCML), Munich, Germany. Correspondence to: Valentyn Melnychuk <melnychuk@lmu.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

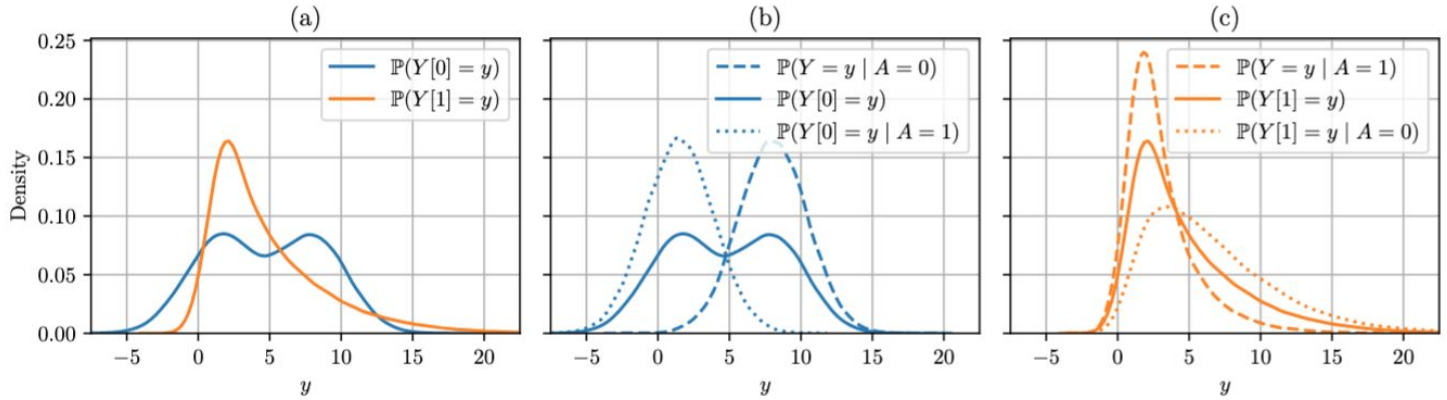
[1]For example, patients with prediabetes are oftentimes treated with metformin monotherapy, which reduces blood glucose sugar (HbA1c) by an *average* of 1.1% (95% confidence interval: 0.9 to 1.3%) (Hirst et al., 2012). Yet, there is often large *skewness* in the potential outcome. While metformin monotherapy is highly effective for some individuals, it fails to achieve glycemic targets for 50% of the patients (Shin, 2019). Here, it is indicated that a second-line anti-diabetes drug is prescribed. Crucially, standard confidence intervals cannot disclose that metformin is harmful to some patients while densities can.

Introduction: Efficient interventional density estimation

Why this is important?

- Making decisions based on averaged causal quantities can be misleading and, in some applications, even dangerous

$$\mathbb{E}(Y[0]) = \mathbb{E}(Y[1]) \approx 4.77$$
$$\text{var}(Y[0]) = \text{var}(Y[1]) \approx 4.06$$



$$\mathbb{P}\{Y[1] < 5.0\} \approx 0.63 \quad \mathbb{P}\{Y[0] < 5.0\} \approx 0.51$$

Problem formulation

Given observational dataset of:

- X covariates
- A treatments
- Y (factual) outcomes

we want to flexibly and efficiently estimate **interventional density** (density of the potential outcomes)

$$\mathbb{P}(Y[a] = y) = \mathop{\mathbb{E}}_{X \sim \mathbb{P}(X)} \big(\mathbb{P}(Y = y \mid X, A = a)\big)$$

Introduction: Task complexity – Assumptions

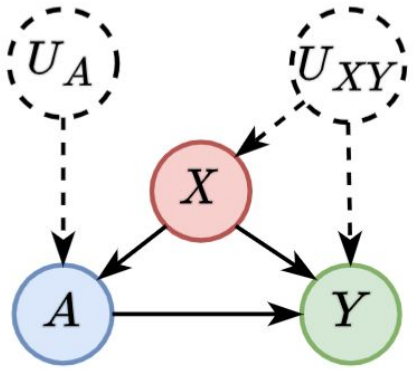
Why estimation is hard?

- Traditional density estimation is non-applicable for **Interventional Density Estimation** (IDE)
- Density is a **functional, infinitely-dimensional target estimand**, and, hence, standard semi-parametric efficiency theory (with influence functions) is not applicable.
- Choice of the nuisance parameters on practice: conditional expectations vs. conditional densities?

Identifiability assumptions

Potential outcomes framework

- **Consistency**. If $A = a$ is a treatment for some patient, then $Y = Y[a]$
- **Positivity (Overlap).** There is always a non-zero probability of receiving/not receiving any treatment, conditioning on the covariates: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$
- **Exchangeability (Ignorability).** Current treatment is independent of the potential outcome, conditioning on the covariates $A \perp\!\!\!\perp Y[a] \mid X$ for all a.



Introduction: Related work - Research gap – Our contributions

Related methods

Method	Parametric	Estimator type	Efficiency wrt.	Base density model	Proper density	Universal
Kim et al. (2018)	semi-parametric	A-IPTW	L_1 distance	kernel density estimation (KDE)	✗	✓
Muandet et al. (2021)	non-parametric	plug-in	—	distributional kernel mean embeddings (DKME)	✗	✓
Kennedy et al. (2023)	semi- / fully-parametric	A-IPTW	moment condition	exponential family	✓	✗
				truncated series (TS)	✗	✓
INFs (this paper)	fully-parametric	A-IPTW	moment condition	normalizing flows (NFs)	✓	✓

A-IPTW: augmented inverse propensity of treatment weighted

Research gap

- Existing methods for IDE are either non- or semi-parametric. Our work is the first to propose a **universal fully-parametric**, deep learning method for IDE, with proper density.

Our contributions

Interventional Normalizing Flows (INFs) are first proper fully-parametric, deep learning method for interventional density estimation:

- We extend the results of (Kennedy et al., 2023) and derive a tractable optimization problem with a one-step bias correction for efficient and doubly robust estimation. This allows for an effective two-step training procedure.
- We demonstrate in various experiments that INFs are highly expressive and effective. A major advantage owed to the parametric form is that our INFs scale well to both large and high-dimensional datasets.

INFs: Semi-parametric IDE (Kennedy et al., 2023)

One-step (semi-parametric) IDE estimators

Target: interventional density.

$$\mathbb{P}(Y[a] = y) = \mathop{\mathbb{E}}_{X \sim \mathbb{P}(X)} \big(\mathbb{P}(Y = y \mid X, A = a)\big)$$

- **Plug-in estimator**:

$$\hat{\mathbb{P}}^{\text{PI}}(Y[a] = y) = \mathbb{P}_n\{\hat{\mathbb{P}}(Y = y \mid X, A = a)\}.$$

Two-step (semi-parametric) IDE estimators

Target: projection parameters

$$T(Y; \beta_a) = -\nabla_{\beta_a} \log g(Y; \beta_a)$$

$$\hat{\beta}_a = \arg\min_{\beta_a} \text{KL}\left(\mathbb{P}(Y[a]) \,\middle\|\, g(\cdot; \beta_a)\right) = \arg\min_{\beta_a} \mathop{\mathbb{E}}_{Y^a \sim \mathbb{P}(Y[a])} \big(-\log g(Y^a; \beta_a)\big).$$

- **Covariate-adjusted estimator:**

$$\hat{\mathbb{P}}^{\text{CA}}(Y[a] = y) = g(y; \hat{\beta}_a^{\text{PI}}) \qquad \hat{m}^{\text{PI}}(\beta_a) = \mathop{\mathbb{E}}_{Y^a \sim \mathbb{P}_n\{\hat{\mathbb{P}}(Y \mid X, A=a)\}} T(Y^a; \beta_a) \stackrel{!}{=} 0$$

- **Augmented inverse propensity of treatment weighted (A-IPTW) estimator:**

$$\hat{\mathbb{P}}^{\text{A-IPTW}}(Y[a] = y) = g(y; \hat{\beta}_a^{\text{A-IPTW}}) \qquad \hat{m}^{\text{A-IPTW}}(\beta_a) = \hat{m}^{\text{PI}}(\beta_a) + \mathbb{P}_n\{\phi_a(T(Y; \beta_a); \hat{\mathbb{P}})\} \stackrel{!}{=} 0.$$

$$\phi_a(T; \mathbb{P}) = \frac{\mathbb{1}(A = a)}{\pi_a(X)}\Big(T - \mathbb{E}(T \mid X, A = a)\Big) + \mathbb{E}(T \mid X, A = a) - \mathop{\mathbb{E}}_{X \sim \mathbb{P}(X)} (\mathbb{E}(T \mid X, A = a)).$$

INFs: Novel efficient optimization objective

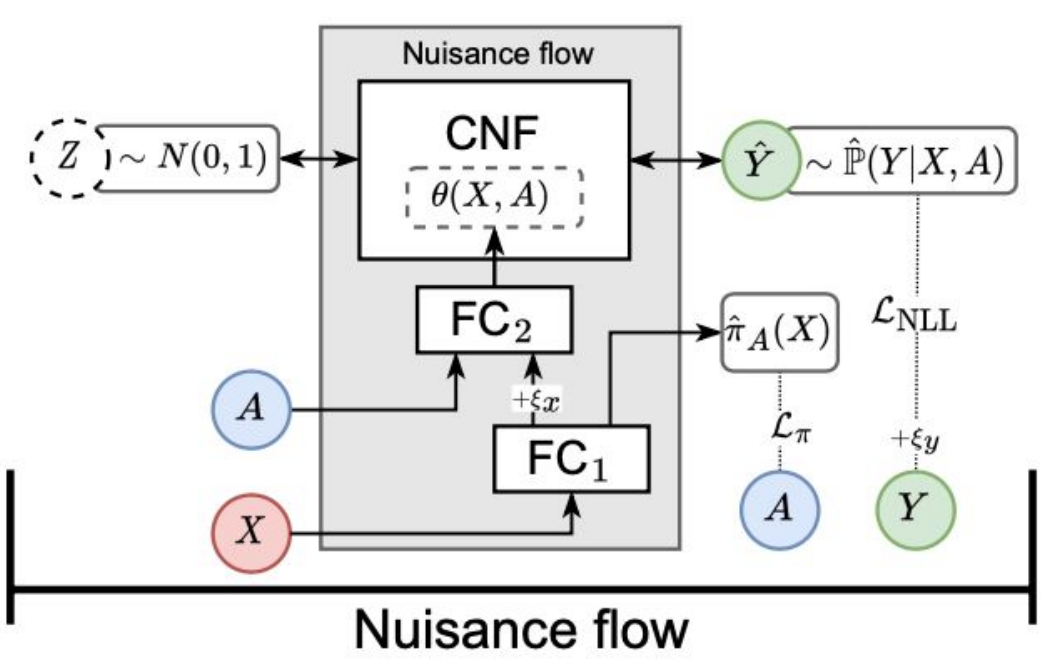
Proposed by (Kennedy et al., 2023)

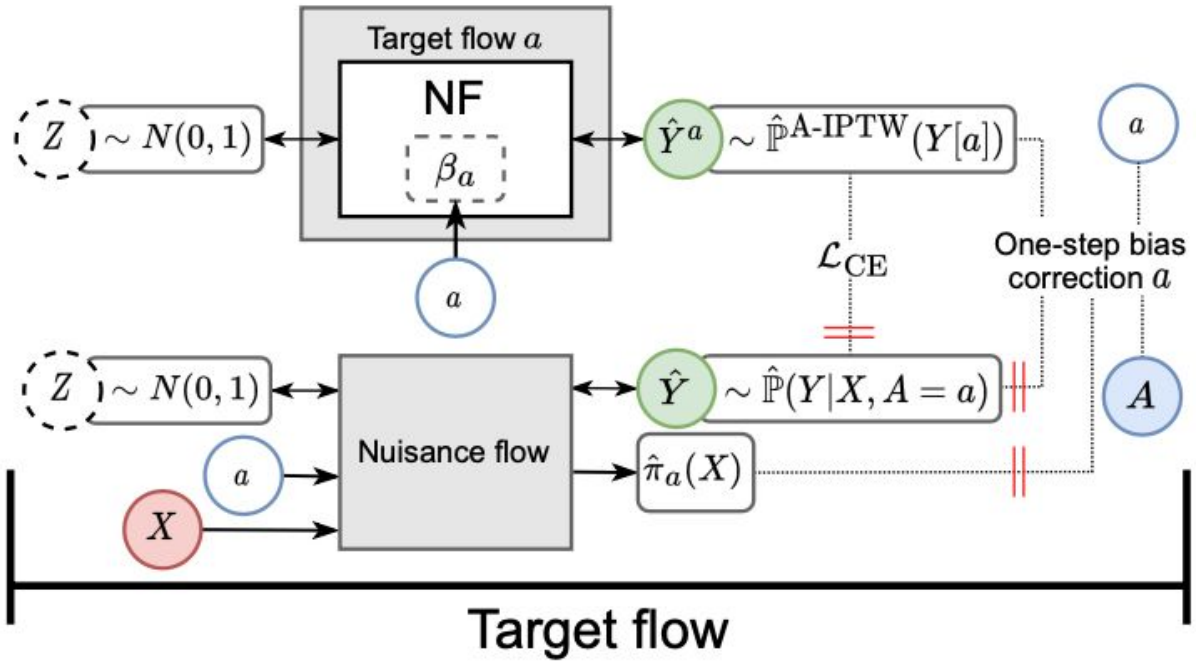
$$\hat{m}^{\text{A-IPTW}}(\beta_a) = \hat{m}^{\text{PI}}(\beta_a) + \mathbb{P}_n\left\{\phi_a(T(Y;\beta_a);\hat{\mathbb{P}})\right\} \stackrel{!}{=} 0$$

Our idea

$$\hat{\beta}_a^{\text{A-IPTW}} = \arg\min_{\beta_a} \Bigg[\underbrace{\mathop{\mathbb{E}}_{Y^a \sim \mathbb{P}_n\{\hat{\mathbb{P}}(Y|X,A=a)\}} \Big(-\log g(Y^a;\beta_a) \Big)}_{\text{cross-entropy loss}} - \underbrace{\mathbb{P}_n\left\{ \frac{\mathbb{1}(A=a)}{\hat{\pi}_a(X)} \Big(\log g(Y;\beta_a) - \mathop{\mathbb{E}}_{Y \sim \hat{\mathbb{P}}(Y|X,A=a)} \big(\log g(Y;\beta_a) \big) \Big) \right\}}_{\text{one-step bias correction}} \Bigg]$$

INFs: Novel architecture





Experiments: Datasets – Results

Datasets

- We evaluate INFs based on 1 synthetic, 77 + 24 + 2 semi-synthetic and 1 real-world datasets
- Only synthetic and semi-synthetic data have ground-truth potential outcomes; real-world evaluation is a proof of concept
- We compared test log-probability for each potential outcome (higher is better)

Results

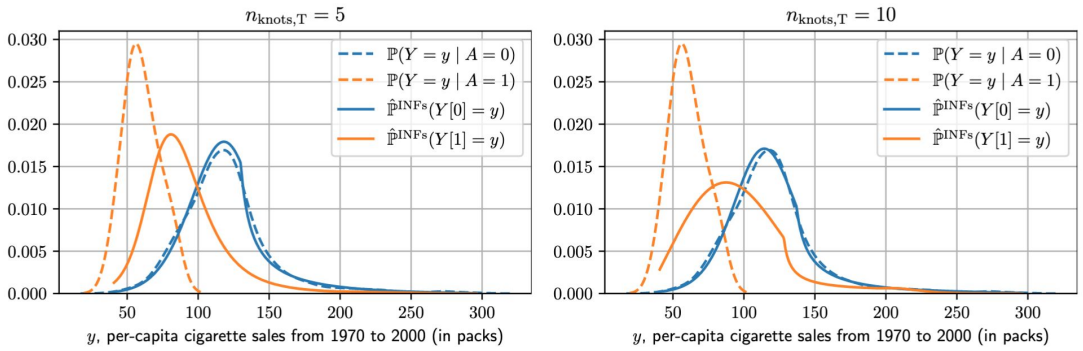
INFs achieve **superior performance** and scales well:

ACIC datasets

	ACIC 2016 (77 datasets)		ACIC 2018 (24 datasets)	
	% best_{in}	% best_{out}	% best_{in}	% best_{out}
TARNet*	3.90%	6.23%	7.08%	7.50%
MDNs	28.96%	29.35%	21.25%	18.75%
CNF [$\widehat{=}$ INFs w/o target flow]	14.42%	15.97%	14.17%	14.58%
KDE (Kim et al., 2018)	1.04%	1.04%	10.42%	9.58%
DKME (Muandet et al., 2021)	0.39%	0.78%	8.75%	10.83%
CNF+TS (Kennedy et al., 2023)	8.18%	8.96%	5.83%	5.42%
INFs w/o bias corr	5.45%	7.27%	4.58%	5.42%
INFs (main)	**37.66%**	**30.39%**	**27.92%**	**27.92%**

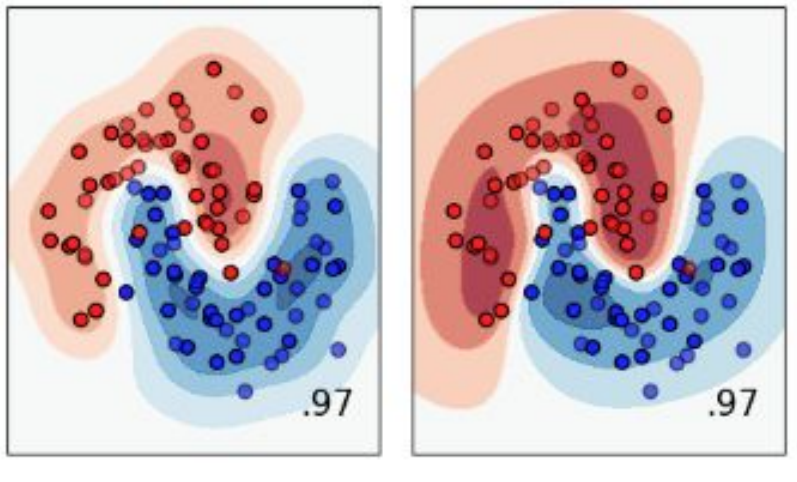
Higher = better (best in bold)

California Tobacco Control Study



Open questions / Future work

- Multi-dimensional outcomes IDE: fundamentally different from the ATE estimation



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

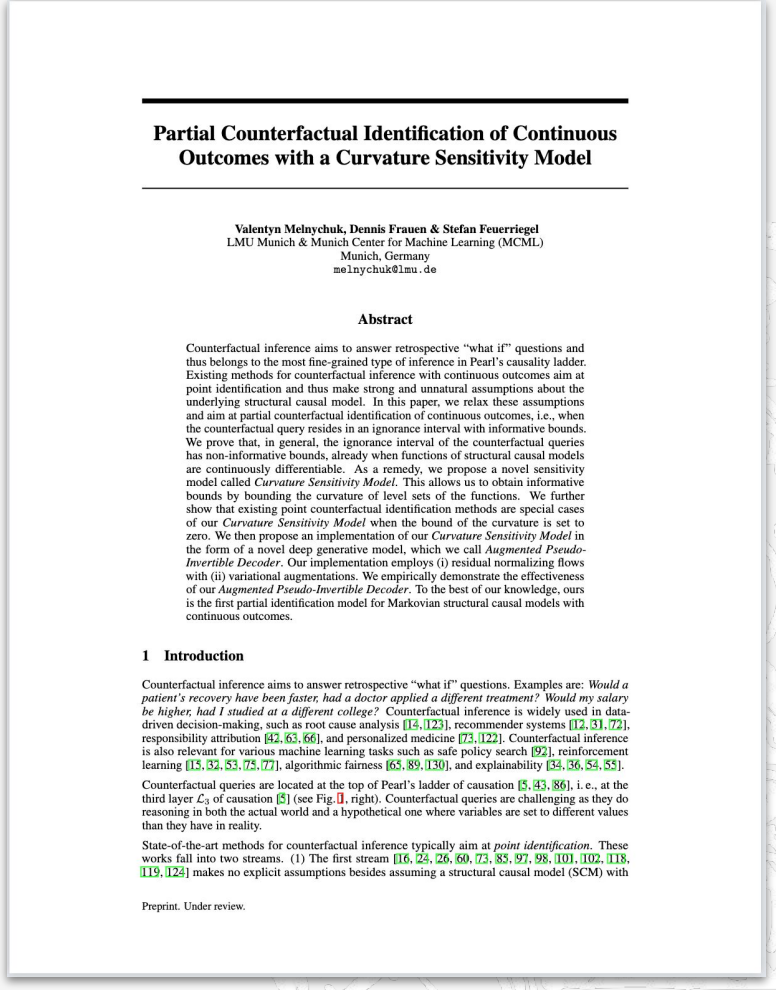
LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

Submitted @ NeurIPS 2023



Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model

Valentyn Melnychuk, Dennis Frauen & Stefan Feuerriegel
LMU Munich & Munich Center for Machine Learning (MCML)
Munich, Germany
melnychuk@lmu.de

Abstract

Counterfactual inference aims to answer retrospective "what if" questions and thus belongs to the most fine-grained type of inference in Pearl's causality ladder. Existing methods for counterfactual inference with continuous outcomes aim at point identification and thus make strong and unnatural assumptions about the underlying structural causal model. In this paper, we relax these assumptions and aim at partial counterfactual identification of continuous outcomes, i.e., when the counterfactual query resides in an ignorance interval with informative bounds. We prove that, in general, the ignorance interval of the counterfactual queries has non-informative bounds, already when functions of structural causal models are continuously differentiable. As a remedy, we propose a novel sensitivity model called *Curvature Sensitivity Model*. This allows us to obtain informative bounds by bounding the curvature of level sets of the functions. We further show that existing point counterfactual identification methods are special cases of our *Curvature Sensitivity Model* when the bound of the curvature is set to zero. We then propose an implementation of our *Curvature Sensitivity Model* in the form of a novel deep generative model, which we call *Augmented Pseudo-Invertible Decoder*. Our implementation employs (i) residual normalizing flows with (ii) variational augmentations. We empirically demonstrate the effectiveness of our *Augmented Pseudo-Invertible Decoder*. To the best of our knowledge, ours is the first partial identification model for Markovian structural causal models with continuous outcomes.

1 Introduction

Counterfactual inference aims to answer retrospective "what if" questions. Examples are: *Would a patient's recovery have been faster, had a doctor applied a different treatment? Would my salary be higher, had I studied at a different college?* Counterfactual inference is widely used in data-driven decision-making, such as root cause analysis [14, 123], recommender systems [12, 31, 72], responsibility attribution [42, 63, 66], and personalized medicine [73, 122]. Counterfactual inference is also relevant for various machine learning tasks such as safe policy search [92], reinforcement learning [13, 32, 53, 75, 77], algorithmic fairness [65, 89, 130], and explainability [34, 36, 54, 55].

Counterfactual queries are located at the top of Pearl's ladder of causation [5, 43, 86], i. e., at the third layer $\mathcal{L}_3$ of causation [5] (see Fig. 1, right). Counterfactual queries are challenging as they do reasoning in both the actual world and a hypothetical one where variables are set to different values than they have in reality.

State-of-the-art methods for counterfactual inference typically aim at *point identification*. These works fall into two streams. (1) The first stream [16, 24, 26, 60, 73, 85, 97, 98, 101, 102, 118, 119, 124] makes no explicit assumptions besides assuming a structural causal model (SCM) with

Preprint. Under review.

Introduction: Counterfactual identification in Markovian SCMs

Why this is important?

- Counterfactual inference is widely used in data-driven decision-making: it aims to answer retrospective "what if" questions
- Counterfactual identifiability is only possible with unnatural or unrealistic assumptions (e.g. monotonicity of the functions in the Markovian SCMs)

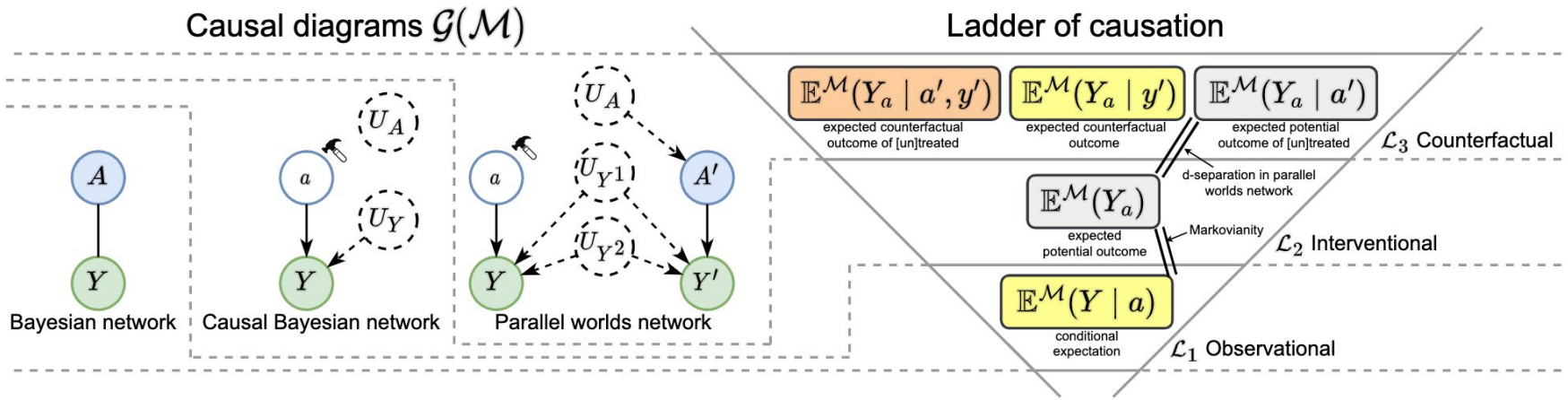
Problem formulation

Given observational dataset of:

- A treatments
- Y (factual) outcomes

we want to perform a partial identification of the **expected counterfactual outcome of [un]treated** ECOU [ECOT]

$$Q^{\mathcal{M}}_{a' \to a}(y') = \mathbb{E}^{\mathcal{M}}(Y_a \mid a', y').$$



Introduction: Task complexity – Related work

Why this is hard?

- Counterfactual queries in general are not identifiable from both L1 and L2 data even for Markovian SCMs.
- Partial identification of L3 discrete outcomes / L2 continuous outcomes does not generalize

Related work

Layer	M/SM	Symbolic identifiability	Point identification methods	Partial identification methods: Discrete outcomes	Partial identification methods: Continuous outcomes
$\mathcal{L}_2$ Interventional	M	Always via back-door criterion [5]	Deep generative models [64, 128]	—	—
	SM	Do-calculus & rules of probability [47, 68, 104]	Potential outcomes framework [10, 23, 114] ; binary IV [33, 49, 115] ; proxy variables [74, 78]	Partially observed back-/front-door variables [70]; canonical SCM [120]	No-assumptions bound [76]; MSM [11, 30, 51, 52, 83, 110]; confounding functions [13, 93]; noisy proxy variables [41]; IV [40, 46, 58, 129]; ATD [3]; clustered DAGs [84]
$\mathcal{L}_3$ Counterfactual	M	Parallel worlds networks [7, 105], counterfactual unnesting theorem [22]	Deep generative models [16, 24, 60, 85, 97, 98, 101, 102] ; Markovian BGMs [50, 56, 79, 80, 107, 132] ; transport-based counterfactuals [27]	PN, PS, PNS [4, 69, 71, 88, 111]; response functions framework / canonical partitioning [4, 82, 96, 121, 125, 126, 127, 131]; causal marginal problem [38, 99]; deep twin networks [113]	**CSM (this paper)**
	SM		ETT [103] ; path-specific effects [106, 130] ; deep generative models [26, 73, 118, 119, 124] ; semi-Markovian BGMs [79]		Future work (see discussion in Appendix E); ANMs with hidden confounding [59]

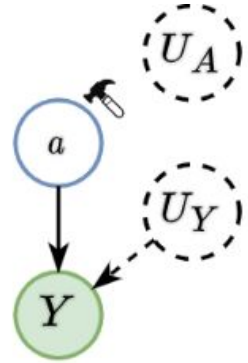
Legend:
- M/SM: Markovian SCM (M), semi-Markovian SCM (SM)

Introduction: Assumptions - Motivating example

Assumptions

- **Bivariate Markovian SCMs** with functions of class C^k and d-dimension latent noise: $\mathfrak{B}(C^k, d)$

$$\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbb{P}(\mathbf{U}), \mathcal{F} \rangle$$
$$\mathbf{U} = \{U_A \in \{0,1\}, U_Y \in [0,1]^d\} \quad \mathcal{F} = \{f_A(U_A), f_Y(A, U_Y)\}$$
$$\mathbf{V} = \{A \in \{0,1\}, Y \in \mathbb{R}\}, \quad \mathbb{P}(\mathbf{U})\text{: } U_A \sim \text{Bern}(p_A),\ 0 < p_A < 1,$$



Motivating example

- ECOU [ECOT] is non-identifiable

Example 1 (Counterfactual non-identifiability in Markovian SCMs). *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two Markovian SCMs from $\mathfrak{B}(C^0, 2)$ with the following functions for Y:*

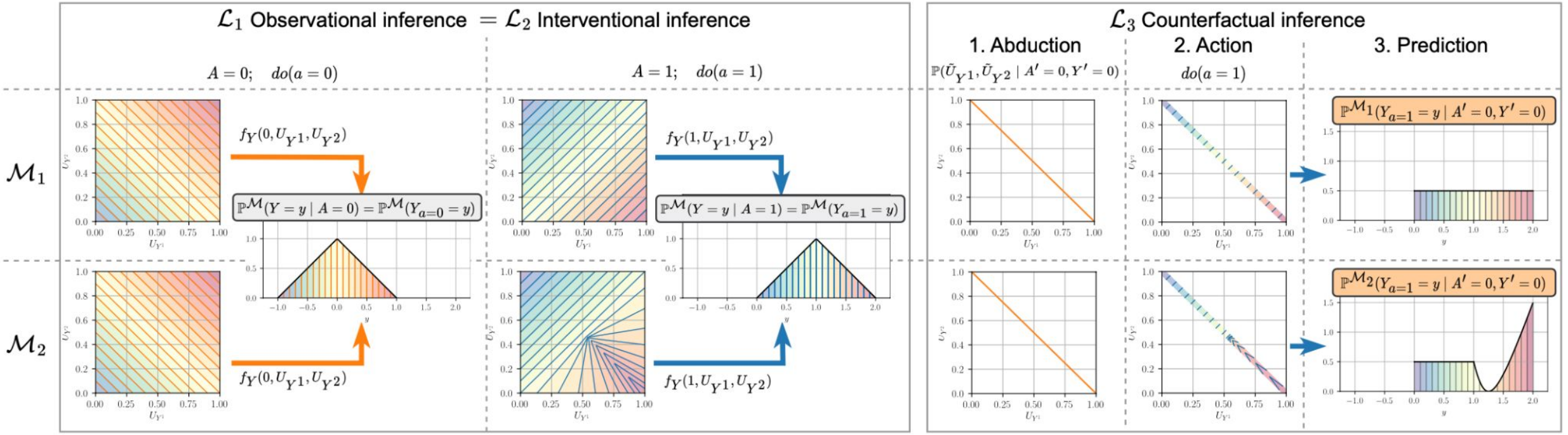
$$\mathcal{M}_1 : f_Y(A, U_{Y^1}, U_{Y^2}) = A\,(U_{Y^1} - U_{Y^2} + 1) + (1 - A)\,(U_{Y^1} + U_{Y^2} - 1)\},$$

$$\mathcal{M}_2 : f_Y(A, U_{Y^1}, U_{Y^2}) = \begin{cases} U_{Y^1} + U_{Y^2} - 1, & A = 0, \\ U_{Y^1} - U_{Y^2} + 1, & A = 1 \wedge (0 \le U_{Y^1} \le 1) \wedge (U_{Y^1} \le U_{Y^2} \le 1), \\ F^{-1}(0, U_{Y^1}, U_{Y^2}), & \textit{otherwise}, \end{cases}$$

where $F^{-1}(0, U_{Y^1}, U_{Y^2})$ is the solution in Y of the implicitly defined function $F(Y, U_{Y^1}, U_{Y^2}) = U_{Y^1} - U_{Y^2} - 2\,(Y-1)\,|-U_{Y^1} - U_{Y^2} + 1| - 1 + \sqrt{(Y-2)^2\,(8\,(Y-1)^2 + 1)} = 0.$

Introduction: Motivating example (continued)

$$Q^{\mathcal{M}_1}_{0\to 1}(0) = 1$$



$$Q^{\mathcal{M}_2}_{0\to 1} \approx 1.114$$

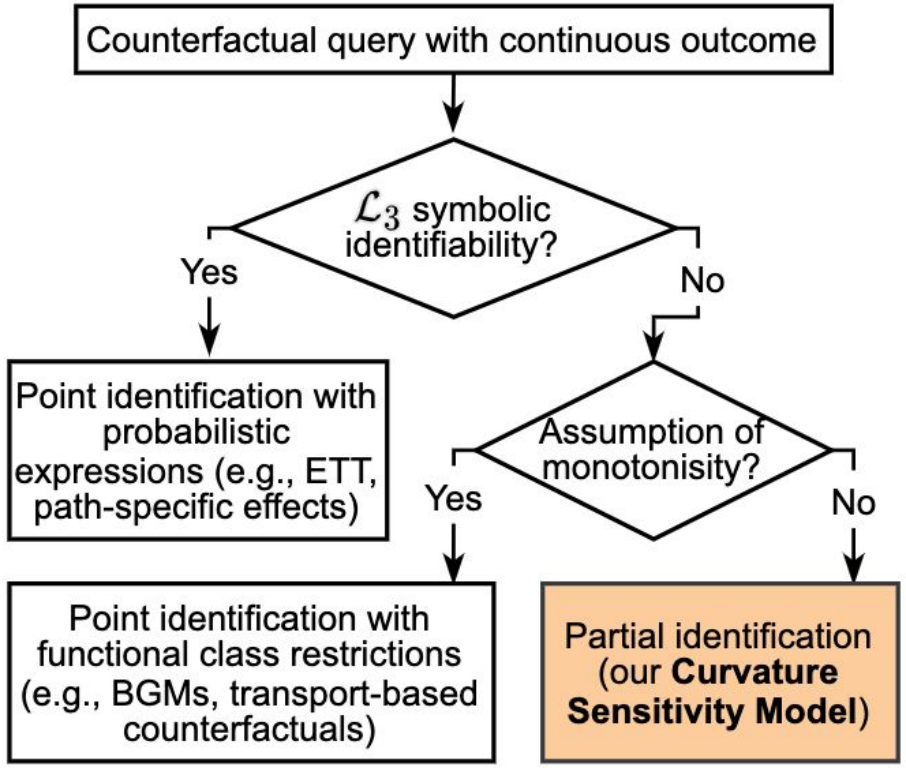
Introduction: Research gap – Our contributions

Research gap

- We are the first to propose a sensitivity model for partial counterfactual identification of continuous outcomes in Markovian SCMs.

Our contributions

- We prove that the expected counterfactual outcome of [un]treated has **non-informative bounds** in the class of continuously differentiable functions of SCMs.
- We propose a novel **Curvature Sensitivity Model (CSM)** to obtain informative bounds. Our CSM is the first sensitivity model for the partial counterfactual identification of continuous outcomes in Markovian SCMs.
- We introduce a novel deep generative model called **Augmented Pseudo-Invertible Decoder (APID)** to perform partial counterfactual inference under our CSM. We further validate it numerically.



Partial Counterfactual Identification - Non-Informative Bounds

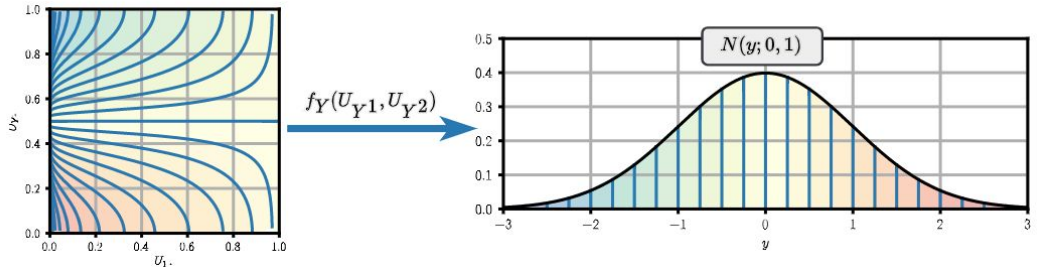
Observational distribution as a pushforward

$$\mathbb{P}^{\mathcal{M}}(Y = y \mid a) = \int_{E(y,a)} \frac{1}{\|\nabla_{u_Y} f_Y(a, u_Y)\|_2} \, \mathrm{d}\mathcal{H}^{d-1}(u_Y)$$

where $E(y, a)$ is a level set (preimage) of y, i. e., $E(y, a) = \{u_Y \in [0, 1]^d : f_Y(a, u_Y) = y\}$, and $\mathcal{H}^{d-1}(u_Y)$ is the Hausdorff measure (see Appendix B for the definition).

Example (Box-Müller transformation)

$$\mathcal{M}_{bm} : f_Y(A, U_{Y^1}, U_{Y^2}) = f_Y(U_{Y^1}, U_{Y^2}) = \sqrt{-2\log(U_{Y^1})}\cos(\pi U_{Y^2})$$



Solution for d=1 (BGMs)

change of variables formula:

$$\mathbb{P}(Y = y \mid a) = \sum_{u_Y \in E(y,a)} |\nabla_{u_Y} f_Y(a, u_Y)|^{-1}$$

\+ monotonicity assumption: $f_Y(a, u_Y) = \mathbb{F}_a^{-1}(\pm u_Y \mp 0.5 + 0.5)$

Partial Counterfactual Identification - Non-Informative Bounds

Counterfactual queries as pushforwards

$$\mathbb{P}^{\mathcal{M}}(Y_a = y \mid a', y') = \frac{1}{\mathbb{P}^{\mathcal{M}}(Y = y' \mid a')} \int_{E(y',a')} \frac{\delta(f_Y(a, u_Y) - y)}{\|\nabla_{u_Y} f_Y(a', u_Y)\|_2} \, \mathrm{d}\mathcal{H}^{d-1}(u_Y),$$

$$Q^{\mathcal{M}}_{a' \to a}(y') = \mathbb{E}^{\mathcal{M}}(Y_a \mid a', y') = \frac{1}{\mathbb{P}^{\mathcal{M}}(Y = y' \mid a')} \int_{E(y',a')} \frac{f_Y(a, u_Y)}{\|\nabla_{u_Y} f_Y(a', u_Y)\|_2} \, \mathrm{d}\mathcal{H}^{d-1}(u_Y)$$

where $E(y', a')$ *is a (factual) level set of* y'*, i. e.,* $E(y', a') = \{u_Y \in [0,1]^d : f_Y(a', u_Y) = y'\}$ *and* $a' \neq a$*.*

\+ monotonicity assumption:

Solution for d=1 (BGMs)

$$Q^{\mathcal{M}}_{a' \to a}(y') = \mathbb{F}_a^{-1}(\pm \mathbb{F}_{a'}(y') \mp 0.5 + 0.5)$$

Partial Counterfactual Identification - Non-Informative Bounds

Partial counterfactual identification of ECOU [ECOT]

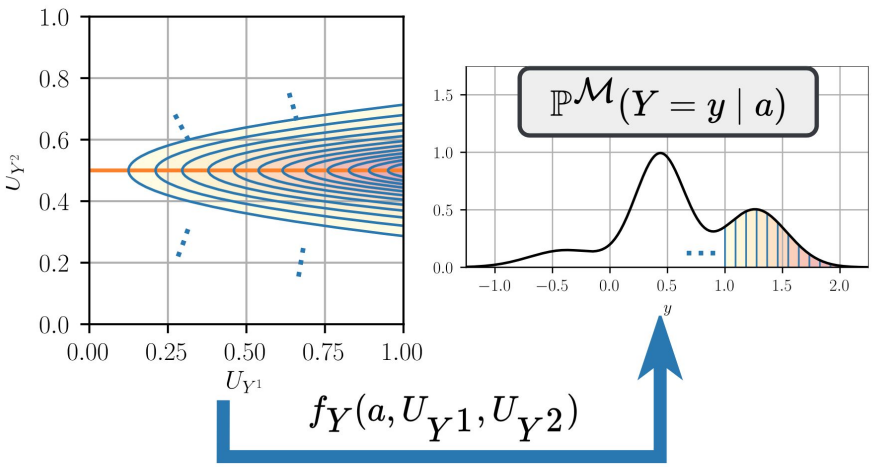
- Constrained variational problem, which involves partial derivatives and Hausdorff integrals:

$$\underline{Q_{a' \to a}(y')} = \inf_{\mathcal{M} \in \mathfrak{B}(C^k, d)} Q^{\mathcal{M}}_{a' \to a}(y') \quad s.t.\ \forall a \in \{0,1\} : \mathbb{P}(Y \mid a) = \mathbb{P}^{\mathcal{M}}(Y \mid a)$$

$$\overline{Q_{a' \to a}(y')} = \sup_{\mathcal{M} \in \mathfrak{B}(C^k, d)} Q^{\mathcal{M}}_{a' \to a}(y') \quad s.t.\ \forall a \in \{0,1\} : \mathbb{P}(Y \mid a) = \mathbb{P}^{\mathcal{M}}(Y \mid a)$$

Non-informative bounds

- **Theorem 1 (informal).** The ignorance interval for the partial identification of the ECOU [ECOT] has **non-informative** bounds for SCMs with functions C^k for every k.

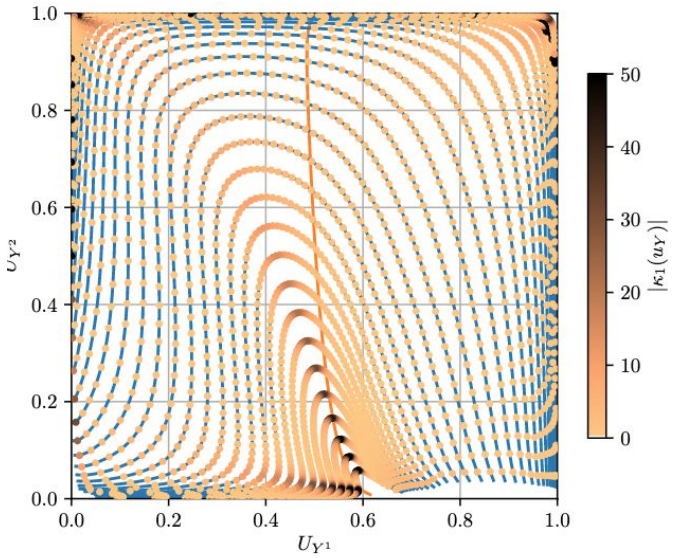


CSM: Assumption Kappa - Informative bounds

Assumption kappa

- (Informal) we assume that κ ≥ 0 is the upper bound of the absolute **curvature** for the level sets.

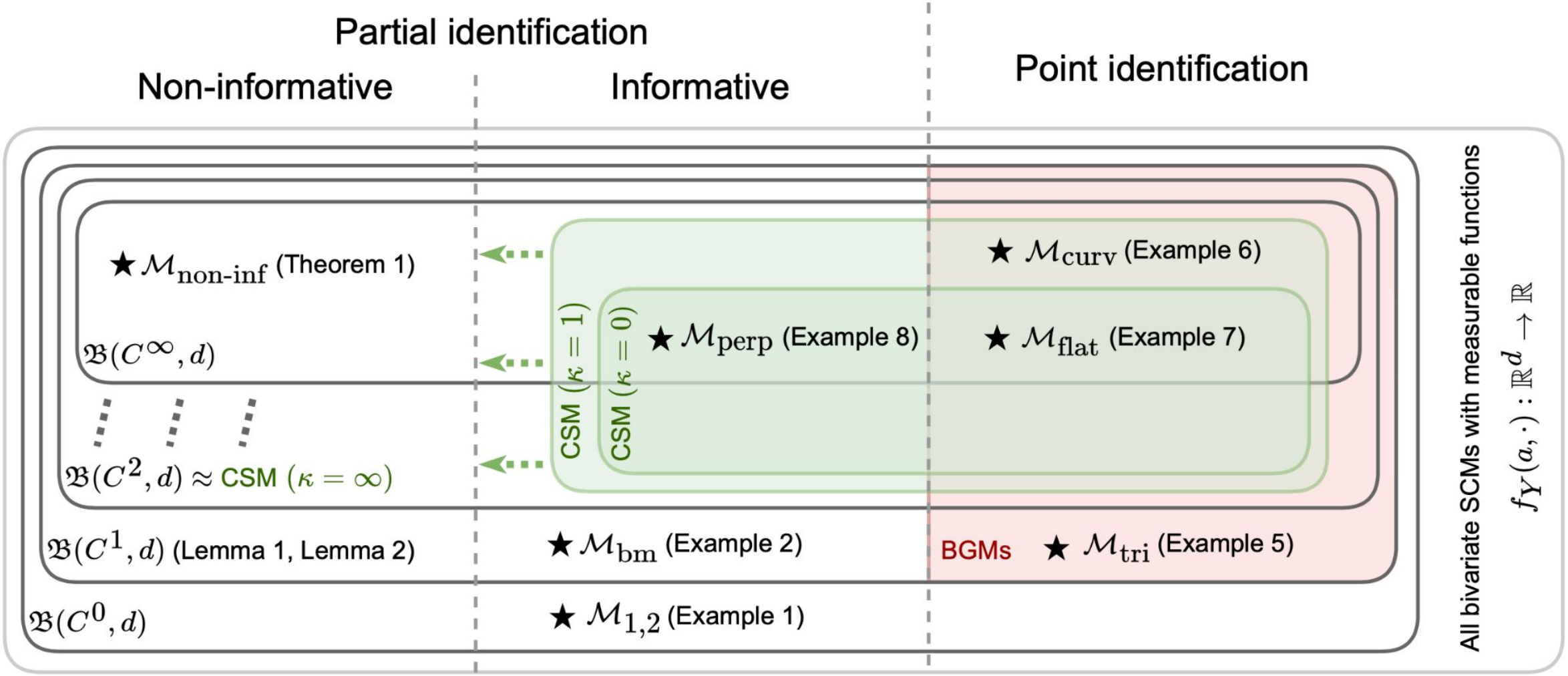
$$\kappa_1(u_Y) = -\frac{1}{2}\nabla_{u_Y}\left(\frac{\nabla_{u_Y} f_Y(a, u_Y)}{\|\nabla_{u_Y} f_Y(a, u_Y)\|_2}\right)$$



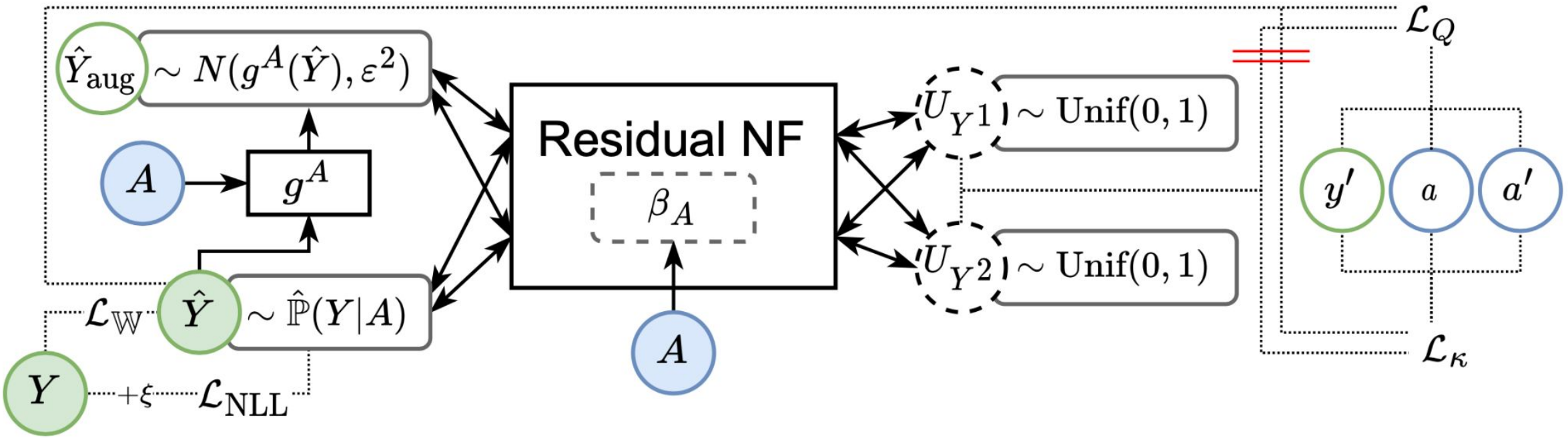
Partial identification with informative bounds

- **Theorem 2 (informal).** Under Assumption kappa, the ignorance interval for the partial identification of the ECOU [ECOT] has **informative** bounds for SCMs with functions C^k for every k > 1.

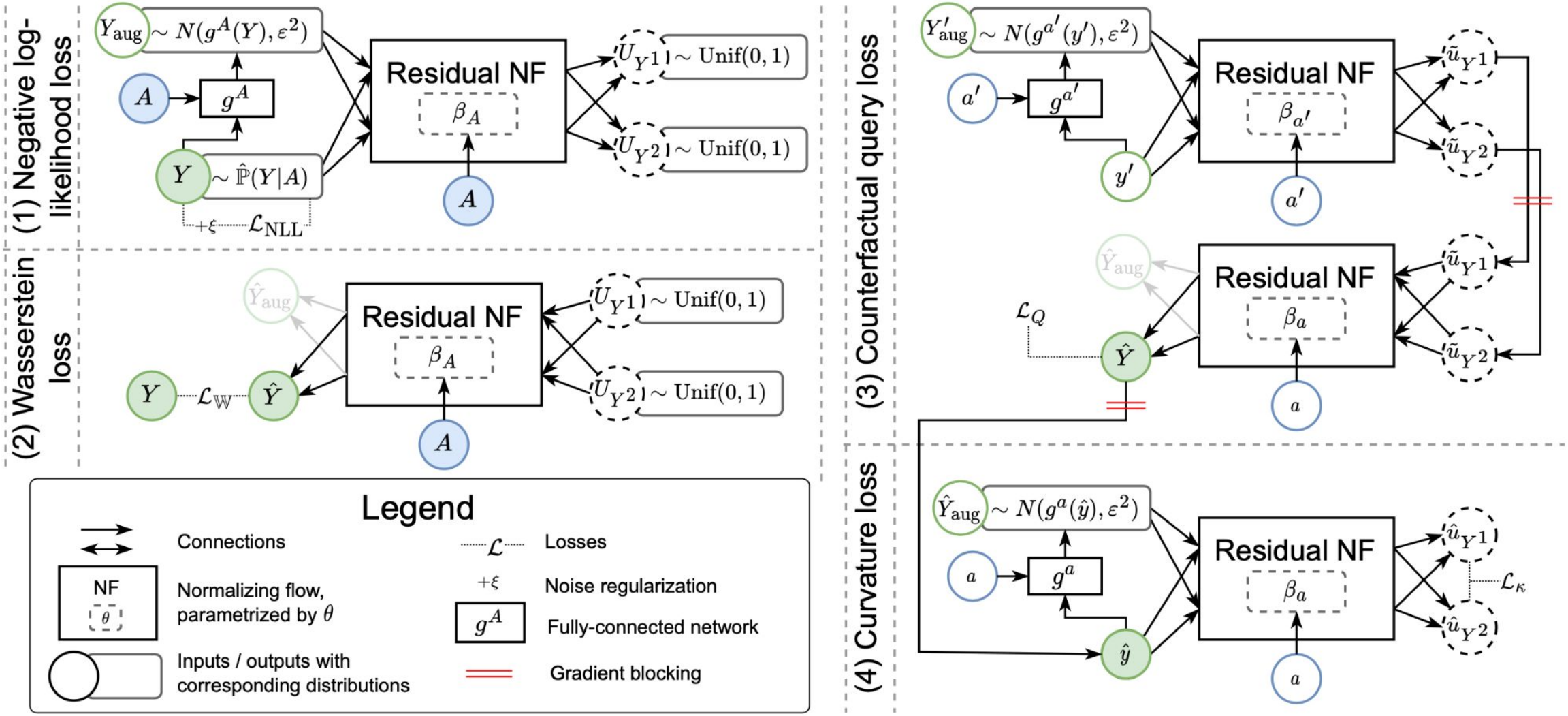
CSM: Identification spectrum



APID: Novel deep generative model



APID: Training



Experiments: Datasets – Results

Datasets

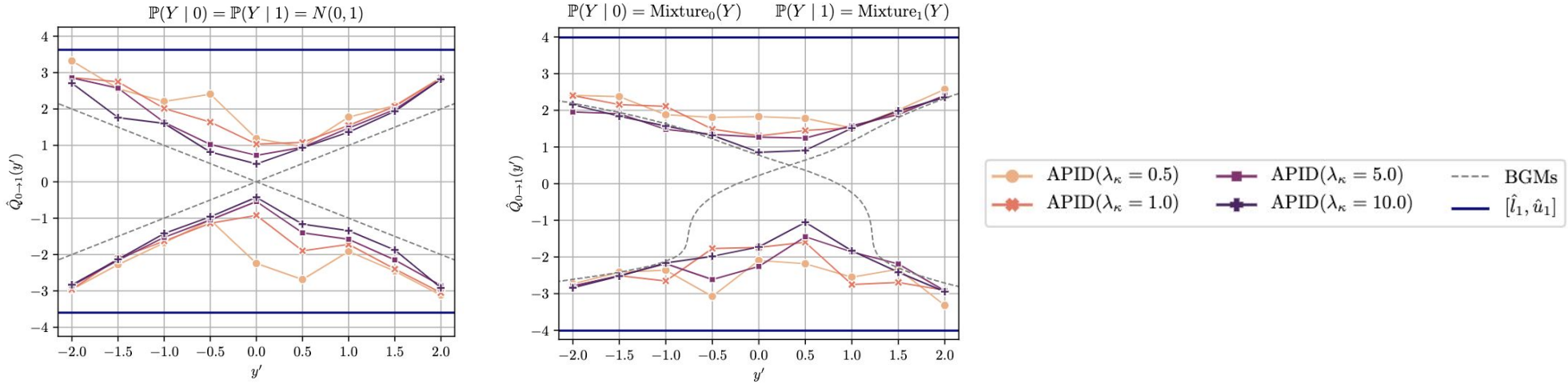
- We evaluate INFs based on 2 synthetic datasets, but even there we do not assume GT SCMs

$$\begin{cases} Y \mid 0 \sim \mathbb{P}(Y \mid 0) = N(0,1) \\ Y \mid 1 \sim \mathbb{P}(Y \mid 1) = N(0,1) \end{cases}$$

$$\begin{cases} Y \mid 0 \sim \mathbb{P}(Y \mid 0) = \mathbf{Mixture}(0.7\,N(-0.5, 1.5^2) + 0.3\,N(1.5, 0.5^2)), \\ Y \mid 1 \sim \mathbb{P}(Y \mid 1) = \mathbf{Mixture}(0.3\,N(-2.5, 0.35^2) + 0.4\,N(0.5, 0.75^2) + 0.3\,N(2.0, 0.5^2)) \end{cases}$$

Results

APID is consistent with BGMs



Open questions / Future work

- More intuition / Connections to a real world
- Combination with Marginal Sensitivity Model for potential outcomes framework (i.e. semi-Markovian SCMs).
- Sharp bounds under CSM (APID does not guarantee tight bounds).